Nayana OCR: A Scalable Framework for Document OCR in Low-Resource Languages

Adithya S Kolavi[1], Samarth P[1], Vyoman Jain[1]

[1]CognitiveLab

Correspondence: adithyaskolavi@gmail.com, samarthprakash8@gmail.com, vyomanjain@gmail.com

Abstract

We introduce Nayana, a scalable and efficient framework for adapting Vision-Language Models (VLMs) to low-resource languages. Despite significant advances, modern VLMs remain constrained by the scarcity of training data in non-English languages, limiting their global applicability. Our framework addresses this fundamental challenge through a novel layout-aware synthetic data generation pipeline combined with parameter-efficient adaptation techniques. Instead of requiring extensive manually annotated datasets, Nayana enables existing models to learn new languages effectively using purely synthetic data. Using Low-Rank Adaptation (LoRA), we demonstrate this capability across ten Indic languages: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, and Telugu. Through extensive experiments in OCR tasks, we show that models can achieve strong performance in new languages without the traditional requirements of large-scale annotated datasets or extensive model modifications. Nayana's success in adapting VLMs to new languages with synthetic data establishes a practical pathway for extending AI capabilities to underserved language communities, particularly in scenarios where annotated data is scarce or unavailable.

1 Introduction

Vision-Language Models (Wang et al. (2024); Wu et al. (2024); Abdin et al. (2024); Chen et al. (2024); Liu et al. (2024a); Wei et al. (2024a)) have demonstrated remarkable success in high-resource languages like English. However, these advancements have not translated across all languages due to a fundamental challenge: the scarcity of high-quality training data. This limitation is particularly evident in languages with complex scripts, where creating large-scale manually annotated datasets is both time-consuming and prohibitively expensive. This has limited the adoption of VLMs for document understanding tasks across diverse languages.

Nayana is an adaptive framework designed to bridge this gap by enabling existing VLMs to learn new languages effectively without requiring extensive annotated datasets. While this paper demonstrates Nayana's capabilities through OCR tasks across ten Indic languages, the framework's approach is inherently flexible and can extend to other tasks and language families. Our methodology eliminates the traditional requirement of annotation by combining synthetic data generation with efficient model adaptation techniques.

The main contributions of this paper are:

1. **Novel Synthetic Data Generation Pipelines:** A layout-aware synthetic data generation pipeline that automates the creation of training datasets while preserving visual and structural relationships in documents. This approach significantly reduces the dependency on manually annotated data for low-resource languages.

2. **Systematic Analysis of LoRA-based Adaptation:** We conduct a comprehensive evaluation of different LoRA techniques and configurations to determine their effectiveness in multilingual adaptation. Our analysis explores whether supervised fine-tuning can enhance language transfer and identifies the optimal configurations for adapting VLMs to new languages with minimal computational overhead.

3. **Comprehensive Empirical Validation:** Through extensive experimentation and evaluation across ten Indic languages, we provide strong evidence that our synthetic data approach matches the performance of traditional OCR Models, establishing a scalable path forward for language adaptation in VLMs.

2 Related Work

Recent Vision Language Models like Qwen 2.5 VL (Wang et al., 2024), Deepseek-VL2 (Wu et al., 2024), InternVL 2.5 (Chen et al., 2024), Llava-NeXT (Liu et al., 2024a), Phi 3.5 Vision (Abdin et al., 2024) have advanced significantly in OCR, captioning, and visual question answering (Antol et al., 2015). These developments stem from parameter-efficient fine-tuning, synthetic data generation, and improved multimodal architectures.

Parameter-efficient fine-tuning methods are crucial for adapting VLMs to specific tasks and languages. Low-Rank Adaptation (Hu et al., 2021) enables efficient parameter updates through low-rank matrix injection in transformer layers.

Multilingual OCR and document understanding have progressed substantially, with systems like Tesseract (Smith, 2007) and PaddleOCR (Du et al., 2020) establishing foundations for multilingual text recognition. Transformer-based approaches like ViLanOCR (Cheema et al., 2024) leverage synthetic data for improved performance on underrepresented languages, while LLaVA-NeXT (Liu et al., 2024a) advances OCR through high-resolution processing and improved visual instruction tuning for training.

Synthetic data generation addresses data scarcity in low-resource settings. SynthVLM (Liu et al., 2024b) uses diffusion models to create image-text pairs, while DocSynth300K (Zhao et al., 2024) demonstrates the effectiveness of generated data for document understanding tasks.

OCR-free approaches offer alternatives to traditional pipelines. DocPedia (Feng et al., 2024) processes documents in the frequency domain, while TextHawk2 (Yu et al., 2024) employs decoder-only architecture with efficient tokenization. Solutions like DocLayout-YOLO (Zhao et al., 2024), Donut (Kim et al., 2021) and Nougat (Blecher et al., 2023) have also explored document understanding without traditional OCR models.

Despite advances in parameter-efficient fine-tuning, synthetic data generation, and OCR-free approaches, challenges persist in adapting VLMs to low-resource languages. Our work introduces language-agnostic synthetic pipelines, combines parameter-efficient tuning with high-resolution vision encoders, and extends OCR-free paradigms to low-resource languages.

3 Synthetic Data Generation: A Scalable Cross-Lingual Framework

The cornerstone of our work lies in developing a sophisticated pipeline for generating high-fidelity synthetic training data that preserves the intricate relationships between document layout, visual elements, and textual content across languages. Our framework addresses the fundamental challenge of data scarcity in low-resource languages through a novel approach that combines advanced document understanding, a state-of-the-art English OCR model, and context-aware translation mechanisms. This section details the architectural components and methodological innovations that enable scalable, high-quality dataset generation for multilingual document understanding tasks.

The pipeline's design emphasizes three critical aspects: preservation of document structure and visual hierarchy, accurate text recognition across diverse scripts, and contextually appropriate translation that maintains semantic integrity. Through careful orchestration of these elements, we achieve a system capable of generating training data that closely mirrors the complexity and nuance of naturally occurring documents while scaling efficiently across multiple languages and document types.

3.1 Seed Dataset Collection

The foundation of our synthetic data generation pipeline rests upon a meticulously curated corpus of English-language documents, encompassing approximately 14,000 distinct samples. Our primary source materials comprise research papers from arXiv (2,000 documents), medical literature from PubMed (1,000 documents), newspaper articles (1,000 pages), and marketing materials (10,000 samples). This collection represents a strategic balance across multiple domains and document types, carefully selected to capture the diverse spectrum of real-world document layouts, content structures and ensures comprehensive coverage of various typographical elements, structural patterns, and domain-specific formatting conventions that characterize modern document ecosystems.

The academic papers, drawn from arXiv's extensive repository, provide exemplars of complex multi-column layouts, mathematical notation, and intricate figure-text relationships. Medical literature from PubMed introduces specialized terminology and standardized reporting formats, while newspaper pages contribute examples of dynamic

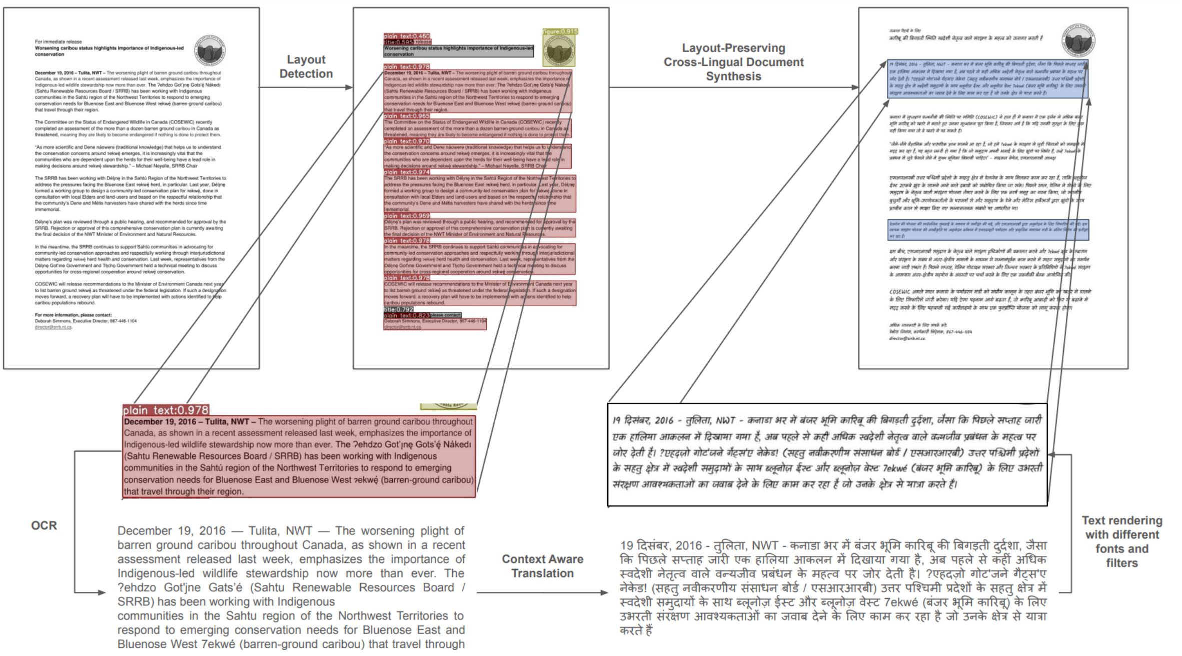


Figure 1: Nayana's end-to-end synthetic data generation pipeline. Starting from English document images, our pipeline generates multilingual datasets for OCR and Document level OCR tasks while preserving layout integrity and visual characteristics. The pipeline processes approximately one image every 3-5 seconds, enabling rapid dataset generation at scale.

layout patterns and diverse content organization. Marketing materials round out the collection with their rich variety of creative layouts, typographical treatments, and visual design elements.

3.2 Multi-stage Processing Pipeline

Our processing methodology employs a sophisticated multi-stage approach that preserves document integrity while enabling efficient multilingual adaptation. The pipeline initiates with high-resolution document preprocessing, converting all inputs to standardized 300 DPI images to ensure consistent quality and feature preservation across source formats. This standardization step establishes a robust foundation for subsequent processing stages.

The layout analysis phase employs an optimized implementation of DocLayout-YOLO (Zhao et al. (2024)), which systematically identifies and classifies document regions including text blocks, titles, figure captions, tables, and visual elements. While our initial research explored ensemble-based approaches using multiple layout detection models, empirical evaluation demonstrated that our optimized single-model implementation achieves comparable accuracy with significantly reduced computational overhead.

Text extraction and visual analysis proceed through a carefully orchestrated sequence of operations. Each identified text region undergoes precise optical character recognition to extract English text from our diverse document collection. We selected Tesseract (Smith (2007)) as the pipeline's OCR model amongst state-of-the-art candidates including PaddleOCR (Du et al. (2020)) and EasyOCR due to its high accuracy at low compute expenditure. The extracted text then undergoes comprehensive visual attribute analysis. This includes background and text color detection, font size estimation, and preservation of critical styling metadata. Our implementation maintains strict fidelity to the original document's visual hierarchy and structural relationships throughout this process.

The translation phase employs a sophisticated multi-engine approach, leveraging several state-of-the-art translation services: Google Translate API, Microsoft Azure Translate, IndicTrans2 (Gala et al. (2023)), and advanced language models such as Llama3.1 405B (Dubey et al. (2024)). This diverse ensemble of translation engines enables robust context-aware translation, with each service contributing its unique strengths in handling different aspects of document context, technical terminology, and formatting conventions.

Our system dynamically selects the most appropriate translation based on context, domain, and

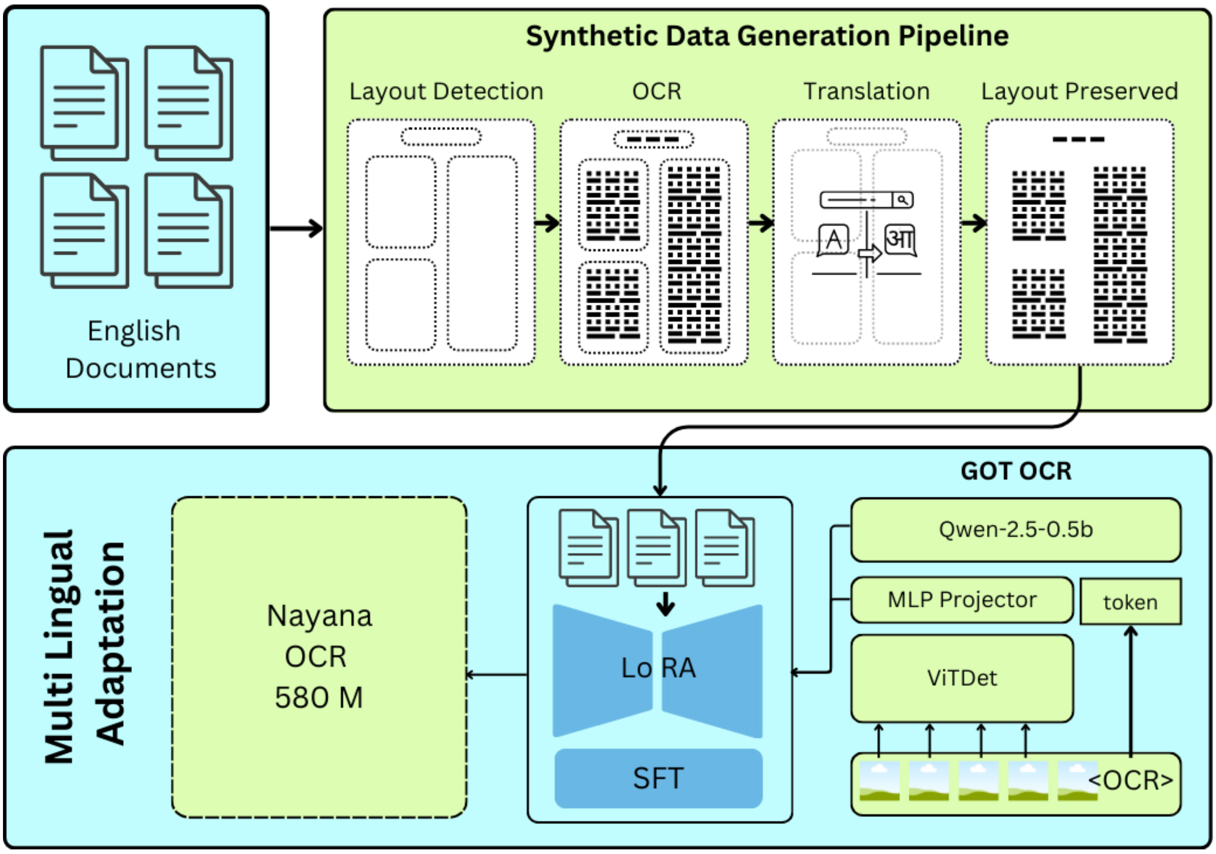


Figure 2: End-to-end Nayana system architecture: (1) A synthetic data generation pipeline transforming English documents into multilingual training data while preserving layout and visual fidelity, (2) OCR model with LoRA adapters for efficient multilingual adaptation, and (3) Training pipeline with Supervised Fine-Tuning (SFT). The modular architecture processes documents in 3-5 seconds while enabling rapid adaptation to new languages with high accuracy.

language pair, ensuring optimal translation quality across diverse document types. The final stage involves precise layout-preserving text replacement (Zhao et al. (2024)), incorporating dynamic font size adjustments and maintaining visual hierarchy while ensuring color contrast preservation.

3.3 Pipeline Performance Characteristics

Our pipeline achieves remarkable efficiency metrics, demonstrating both speed and accuracy at scale. Processing individual documents in approximately 3-5 seconds, the system maintains exceptional performance across all processing stages while enabling rapid dataset generation for new languages. The optimized DocLayout-YOLO (Zhao et al. (2024)) implementation consistently achieves 95.8% accuracy in structural analysis, while the OCR model and sophisticated translation architecture work in concert to ensure high-quality text extraction and translation.

The system's effectiveness is particularly evident in its data multiplication capabilities. Through our augmentation strategies and multi-task generation approach where we use the layout data to extract region-specific information, we achieve an output multiplication factor of 7-10× images per source document. This rich extraction process yields diverse training signals including layout structures, text content. The extracted multi-modal elements can be leveraged for training various downstream models such as VLMs for Visual Question Answering (Antol et al., 2015), Information Extraction systems, Multi-Modal Retrievers(Faysse et al., 2025).This multiplication effect significantly amplifies the utility of our seed dataset, enabling the creation of comprehensive training sets from a relatively modest collection of source documents. The combination of speed, multiplication factor, and rich multi-modal data extraction makes our pipeline particularly effective for rapidly bootstrapping vision-language capabilities in new languages and diverse document understanding applications.

4 Architectural Innovation: Parameter-Efficient Cross-Script Learning

The adaptation of vision-language models (VLMs) for multilingual document understanding presents a fundamental architectural challenge: How to effectively extend models trained primarily on Latin scripts to handle dramatically different writing systems while maintaining computational efficiency. This section details our systematic exploration of architectural approaches, empirically-driven design decisions, and the development of our parameter-efficient adaptation methodology.

Our initial investigation began with a comprehensive evaluation of contemporary VLM architectures, analyzing their fundamental capabilities in handling text-dense images. This exploration revealed a critical insight: while many models excel at general visual understanding, they often struggle with the precise geometric and spatial relationships inherent in document processing. Through extensive experimentation with architectures ranging from traditional CNN-based models to state-of-the-art transformer variants, we identified several key architectural requirements that would prove crucial for successful cross-script adaptation.

4.1 Foundation Model Selection and Analysis

The selection of an appropriate foundation model emerged from a rigorous empirical study evaluating multiple state-of-the-art architectures. Our investigation focused particularly on models' ability to handle the unique challenges presented by Indic scripts, including complex ligatures, overlapping characters, and varied writing directions. Initial experiments with popular vision-language models revealed significant limitations in handling dense textual content, despite their strong performance on general vision-language tasks.

The breakthrough came through our analysis of GOT OCR (580M parameters) (Wei et al. (2024b)), which demonstrated exceptional performance across key metrics. Based on published benchmarks, GOT OCR achieved superior results with an Edit Distance of 0.035/0.038 and F1-scores of 0.972/0.980 for English and Chinese respectively, significantly outperforming larger models like Qwen-VL-Max (>72B parameters) (Wang et al. (2024)) and Vary (7B parameters) (Wei et al. (2024a)). More importantly, its architecture demonstrated remarkable flexibility in handling non-Latin scripts, likely due to its original design for handling both English and Chinese characters – writing systems with significantly different visual characteristics.

Our choice of GOT OCR (Wei et al. (2024b)) was further validated through its optimal balance of performance and efficiency due to its:

- Superior vision transformer backbone architecture compared to contemporary VLM designs
- Specialized text detection heads optimized for dense textual content
- Efficient parameter count (580M) enabling practical deployment while maintaining state-of-the-art performance

The model's architecture, particularly its attention mechanisms and hierarchical feature processing, provided an ideal foundation for our cross-script adaptation strategy. Notably, its transformer-based design facilitated efficient parameter adaptation through Low-Rank Adaptation (Hu et al. (2021)), enabling us to preserve the model's fundamental visual understanding while extending its capabilities to new scripts.

During our initial exploration phase, we pursued several alternative approaches that, while ultimately unsuccessful, provided crucial insights. We conducted extensive experiments with vocabulary expansion techniques, hypothesizing that direct modification of the tokenization layer would enable better handling of Indic scripts. These experiments involved:

- Direct vocabulary expansion with script-specific tokens
- Hierarchical tokenization schemes for handling complex ligatures
- Script-aware embedding layer modifications

Despite systematic exploration of these approaches with various hyperparameter configurations, the results consistently plateaued at 50-60% accuracy for both training and evaluation. This empirical evidence led us to a crucial realization: the challenge lay not in the vocabulary representation but in the fundamental visual processing of different scripts.

4.2 Cross-Modal Alignment Learning

The Cross-Modal Alignment (CMA) phase extends GOT OCR's (Wei et al., 2024b) capabilities beyond its original English and Chinese training domain through a two-phase training approach. Built on GOT OCR's task-token architecture (e.g., <OCR>), our adaptation strategy systematically builds multilingual capabilities while preserving the model's core strengths.

The first phase focuses on section-level training 15, where we use layout-preserving translation to create training pairs from dense textual sections. By unfreezing all major components (ViTDet vision encoder, MLP projection layer, and Qwen 0.5B language model), we enable comprehensive adaptation to new language patterns. Ablation studies confirmed this phase's criticality - attempts to skip directly to document-level training resulted in stalled learning and hallucinations.

The second phase transitions to document-level OCR 10, training on complete document images while selectively freezing components. We maintain the trained visual features by freezing the ViTDet vision encoder while continuing to train the language model and projection layer. This approach successfully extends the model's capabilities to new languages while preserving its performance on English and Chinese texts.

Table 1: Training Phase Configuration Summary

Component	Phase 1	Phase 2
ViTDet Vision Encoder	Unfrozen	Frozen
MLP Projection Layer	Unfrozen	Unfrozen
Qwen 0.5B LLM	Unfrozen	Unfrozen
Training Data	Text-heavy Sections	Complete Documents

4.3 Single-Language Adaptation Results

4.3.1 Hindi Adaptation Performance

Our initial experiments with Hindi adaptation revealed several crucial insights about parameter-efficient adaptation strategies. The choice of 85,000 image-text pairs was determined through extensive preliminary testing, which showed that this dataset size provided optimal coverage of Hindi script variations while remaining computationally manageable.

The results in Table 2 demonstrate a clear progression in adaptation effectiveness across different configurations. The baseline LoRA configuration (r=32, α =64) established fundamental script adaptation but showed limitations in handling complex Hindi character combinations, as evidenced by its BLEU score of 0.29. The optimal configuration (r=64, α=128) achieved substantially better performance, with a BLEU score of 0.58, through improved capacity for modeling intricate script-specific features.

Particularly noteworthy is the preservation of English language capabilities. While the higher-rank LoRA configuration showed a slight decrease in English BLEU scores (from 0.84 to 0.79), it maintained strong overall performance (F1: 0.86, METEOR: 0.88), suggesting effective balance between adaptation and preservation of base capabilities.

Table 2: Hindi Adaptation Performance Comparison

Configuration	Lang	BLEU↑	ANLS↑	F1↑	METEOR↑
LoRA	Hindi	0.29	0.71	0.56	0.57
(r=32, α=64)	English	0.84	0.97	0.91	0.91
LoRA	Hindi	**0.58**	**0.91**	**0.76**	**0.77**
(r=64, α=128)	English	0.79	0.97	0.86	0.88
Full	Hindi	0.50	0.86	0.75	0.73
Fine-tune	English	0.74	0.95	0.85	0.85

4.3.2 Tamil Adaptation Performance

The Tamil adaptation experiments presented unique challenges due to the script's distinctive characteristics, including its cursive nature and complex grapheme structure. Table 3 reveals several important patterns in adaptation behavior. The LoRA configuration (r=64, α=128) demonstrated remarkable robustness in handling Tamil's unique script features, achieving a BLEU score of 0.37 despite the script's significant divergence from the model's original training domain. This performance is particularly impressive given Tamil's complex vowel modification system and the presence of compound characters that can span multiple positions. The comparison with full fine-tuning is especially illuminating. While full fine-tuning achieved reasonable performance (ANLS: 0.79), it showed significant degradation in English capabilities, suggesting potential catastrophic forgetting. In contrast, our LoRA approach maintained strong performance across both languages, with English metrics remaining notably stable (BLEU: 0.78, F1: 0.87).

4.4 Multi-Language Adaptation

We investigated three distinct approaches to handling multiple scripts simultaneously, each offering unique insights into cross-lingual transfer. The Single LoRA approach emerged as particularly effective, demonstrating strong performance across

Table 3: Tamil Adaptation Performance Comparison

Configuration	Lang	BLEU↑	ANLS↑	F1↑	METEOR↑
LoRA	Tamil	**0.37**	**0.87**	**0.66**	**0.64**
(r=64, α=128)	English	**0.78**	**0.96**	**0.87**	**0.88**
Full	Tamil	0.17	0.79	0.44	0.44
Fine-tune	English	0.69	0.96	0.76	0.80

multiple languages without requiring explicit language specification during inference. When language tags were provided both during training and inference, we observed further improvements in performance. A notable advantage of this approach was its ability to leverage cross-script learning - for instance, the model showed improved handling of Marathi text despite being primarily trained on Hindi, suggesting effective transfer between related Devanagari scripts. The Multi-LoRA approach, training separate LoRA modules for each language, achieved strong language-specific performance but sacrificed the beneficial cross-script transfer effects observed in the single LoRA strategy. Despite its strong per-language performance, this approach's inability to leverage script similarities represented a significant limitation in the multilingual context. Nayana We also explored a Merged LoRA strategy, where independently trained language-specific LoRAs were combined using model merging techniques. While this approach showed promising results for both languages, it did not outperform the single LoRA approach's ability to capture cross-script features.

Table 4: Multi-Language Adaptation Performance (Hindi + Kannada) in a single LoRA

Configuration	Lang	BLEU↑	ANLS↑	F1↑	METEOR↑
Single	Hindi	0.64	0.89	0.85	0.84
LoRA	Kannada	0.52	0.72	0.55	0.43
	English	0.79	0.97	0.86	0.88

5 Results

5.1 Evaluation Methodology

Our evaluation framework was designed to provide rigorous, comprehensive assessment across diverse document types and writing systems. We constructed a carefully balanced test set comprising 500 images per language, strategically distributed across different document categories to ensure broad coverage of real-world scenarios. The dataset draws 40% of its content from academic papers sourced from arXiv, another 40% from medical literature in PubMed, and the remaining 20% split equally between newspaper content and advertising materials. This distribution reflects the varying complexity and specialized requirements of different document processing applications.

To ensure methodological rigor and fair cross-linguistic comparison, we developed parallel versions of each document across all ten languages while maintaining identical visual layouts and content structures. This parallel corpus approach enables precise isolation of script-specific challenges while controlling for variations in document complexity and formatting. Such controlled comparison proves essential for understanding the true impact of script differences on model performance.

5.2 Comparative Analysis

Our comprehensive evaluation framework encompasses three distinct categories of document processing systems, each representing different approaches to multilingual document understanding. We first examined traditional OCR systems, including industry standards like Tesseract Smith (2007) and PaddleOCR (Du et al. (2020)), which have established strong baselines in multilingual text recognition. These systems, while specialized for OCR tasks, provide important reference points for performance evaluation.

The second category comprises recent vision-language models, including cutting-edge systems like Phi-3.5 Vision (Abdin et al. (2024)) and Llama-3.2 (Dubey et al. (2024)). These models, despite their impressive capabilities in general vision-language tasks, demonstrate the ongoing challenges in specialized document processing. Our analysis of their performance reveals important insights about the limitations of general-purpose architectures when applied to script-specific document understanding tasks.

Our Nayana-OCR variants, built upon the GOT OCR (Wei et al. (2024b)) architecture, represent the third category. Through extensive training on approximately 850,000 synthetic images spanning 10 Indic languages, these models demonstrate significant advantages in multilingual document processing. The results reveal substantial improvements across key metrics, most notably a 76% reduction in Character Error Rate compared to the base GOT OCR model. This improvement is particularly significant given that it maintains consistency across all evaluated languages.

The performance gains extend beyond simple

Table 5: Average performance metrics across all evaluated languages. Results show mean values for each model across the ten tested languages. Lower values (↓) are better for CER and WER, while higher values (↑) are better for other metrics. Best results in each category are highlighted in **bold**.

Model	CER↓	WER↓	BLEU↑	ANLS↑	METEOR↑
Tesseract	**0.206**	0.583	0.318	**0.797**	0.540
PaddleOCR	0.621	0.880	0.020	0.287	0.069
Llama-3.2 11B	3.858	3.900	0.007	0.091	0.055
Phi-3.5 Vision	2.420	2.461	0.007	0.086	0.044
Qwen2-VL 2B	1.776	1.793	0.025	0.129	0.086
GOT-OCR	0.945	1.041	0.016	0.071	0.052
Nayana-OCR	0.227	**0.463**	**0.395**	**0.796**	**0.630**

character recognition. Our models show markedly improved BLEU scores, indicating enhanced capability in handling complex linguistic structures and maintaining semantic coherence. The reduced standard deviations across performance metrics suggest robust cross-language stability, a crucial factor for practical deployment in multilingual environments. These improvements stem from our innovative approach to model adaptation and the sophisticated synthetic data generation pipeline described in previous sections.

5.3 Detailed Performance Analysis

Examining Table 6, several patterns emerge that illustrate the strengths and limitations of different approaches. Traditional OCR systems like Tesseract (Smith (2007)) show strong performance in character-level accuracy (CER: 0.206) but struggle with higher-level semantic understanding, as evidenced by lower BLEU scores (0.318). In contrast, Nayana-OCR achieves competitive character-level accuracy (CER: 0.227) while substantially outperforming all baselines in semantic metrics (BLEU: 0.395).

The performance gap between general-purpose vision-language models and specialized OCR systems is particularly noteworthy. Despite their larger parameter counts, models like Llama-3.2 11B (Dubey et al. (2024)) and Phi-3.5 Vision (Abdin et al. (2024)) show significantly higher error rates across all metrics. This disparity underscores the importance of architectural choices specifically optimized for document understanding tasks.

5.4 Limitations and Future Work

While our approach demonstrates significant progress, several limitations should be noted. When compared to traditional OCR systems, our models show higher inference latency, reflecting the complexity of vision-language processing. Performance variations across scripts suggest room for improvement in handling certain complex writing systems. Additionally, our synthetic data generation, while efficient, may not capture all real-world variations in document layouts and styles.

Future work will focus on expanding the diversity of seed datasets, incorporating more complex document structures, and developing specialized architectures that better balance performance and computational efficiency. We also plan to explore how our synthetic data approach can benefit other vision-language tasks and create open-source tools to facilitate broader adoption of multilingual vision-language technologies.

6 Conclusion

This work establishes that vision-language models can be effectively adapted to new languages using purely synthetic data, reducing dependency on costly manual annotation. Our results demonstrate that Nayana provides a practical, scalable solution for extending AI capabilities to low-resource languages. By achieving strong performance across diverse scripts while maintaining computational efficiency, our framework opens new possibilities for democratizing AI technologies across linguistic boundaries. The success of our approach not only validates the effectiveness of synthetic data generation and efficient adaptation techniques but also establishes a promising direction for developing more inclusive AI systems that can serve diverse linguistic communities worldwide.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.

Musa Dildar Ahmed Cheema, Mohammad Daniyal Shaiq, Farhaan Mirza, Ali Kamal, and M Asif Naeem. 2024. Adapting multilingual vision language transformers for low-resource urdu optical character recognition (ocr). *PeerJ Computer Science*, 10:e1964.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

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

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*, 67(12):1–14.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7(15):2.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Zheng Liu, Hao Liang, Xijie Huang, Wentao Xiong, Qinhan Yu, Linzhuang Sun, Chong Chen, Conghui He, Bin Cui, and Wentao Zhang. 2024b. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. *arXiv preprint arXiv:2407.20756*.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 629–633, Washington, DC, USA. IEEE Computer Society.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. 2024b. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Ya-Qi Yu, Minghui Liao, Jiwen Zhang, and Jihao Wu. 2024. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024. Doclayout-yolo: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *arXiv preprint arXiv:2410.12628*.

A Appendix

A.1 Language-wise Performance Analysis

Table 6: Detailed Performance Analysis Across Languages. The table compares various OCR models across multiple languages using metrics such as CER, WER, BLEU, ANLS and METEOR.

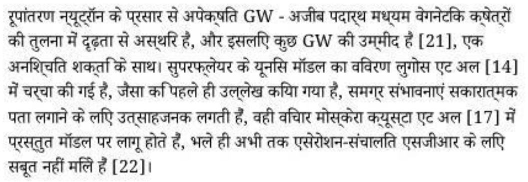
Model	CER↓	WER↓	BLEU↑	ANLS↑	METEOR↑
Hindi					
Tesseract	**0.090**	**0.287**	**0.636**	**0.908**	**0.791**
PaddleOCR	0.414	0.864	0.023	0.575	0.117
Phi-3.5 Vision	2.878	2.500	0.023	0.126	0.069
Llama-3.2 11B	4.654	3.455	0.020	0.116	0.070
Qwen2-VL 2B	2.360	2.022	0.066	0.172	0.153
GOT OCR base	1.013	1.190	0.004	0.052	0.043
Nayana-OCR	0.160	0.297	0.532	0.850	0.756
Kannada					
Tesseract	**0.155**	**0.609**	0.259	**0.847**	0.541
PaddleOCR	0.814	0.918	0.020	0.110	0.048
Phi-3.5 Vision	2.655	2.877	0.006	0.084	0.046
Llama-3.2 11B	4.670	4.991	0.004	0.075	0.047
Qwen2-VL 2B	1.394	1.599	0.013	0.075	0.063
GOT OCR base	0.936	1.008	0.019	0.067	0.063
Nayana-OCR	0.361	0.648	**0.341**	0.740	**0.554**
Tamil					
Tesseract	0.265	0.811	0.109	0.750	0.324
PaddleOCR	0.545	1.076	0.003	0.450	0.051
Phi-3.5 Vision	1.531	2.033	0.000	0.082	0.035
Llama-3.2 11B	3.009	4.229	0.002	0.086	0.052
Qwen2-VL 2B	1.260	1.515	0.007	0.125	0.053
GOT OCR base	0.956	1.020	0.013	0.056	0.051
Nayana-OCR	**0.181**	**0.551**	**0.377**	**0.829**	**0.592**
Telugu					
Tesseract	**0.158**	0.589	**0.296**	**0.821**	**0.551**
PaddleOCR	0.435	0.934	0.014	0.550	0.088
Phi-3.5 Vision	2.442	2.464	0.001	0.067	0.036
Llama-3.2 11B	2.736	3.586	0.015	0.090	0.068
Qwen2-VL 2B	1.580	1.696	0.010	0.115	0.065
GOT OCR base	0.925	1.007	0.022	0.075	0.066
Nayana-OCR	0.282	**0.065**	0.241	0.733	0.522
Odia					
Tesseract	**0.290**	0.681	0.155	0.703	0.403
PaddleOCR	0.639	**0.742**	0.020	0.111	0.030
Phi-3.5 Vision	2.311	2.168	0.000	0.090	0.018
Llama-3.2 11B	2.880	2.908	0.005	0.088	0.042
Qwen2-VL 2B	1.247	1.345	0.012	0.092	0.060
GOT OCR base	0.926	1.000	0.020	0.078	0.042
Nayana-OCR	0.311	0.566	**0.305**	**0.738**	**0.551**

Model	CER↓	WER↓	BLEU↑	ANLS↑	METEOR↑
Punjabi					
Tesseract	0.203	0.532	0.356	0.803	0.568
PaddleOCR	0.717	0.811	0.010	0.095	0.029
Phi-3.5 Vision	3.431	2.896	0.001	0.083	0.034
Llama-3.2 11B	5.801	4.535	0.000	0.065	0.029
Qwen2-VL 2B	1.260	1.515	0.007	0.125	0.053
GOT OCR base	0.954	0.994	0.010	0.066	0.046
Nayana-OCR	**0.159**	**0.440**	**0.435**	**0.853**	**0.693**
Malayalam					
Tesseract	0.355	0.828	0.065	0.663	0.258
PaddleOCR	0.788	**0.895**	0.036	0.125	0.073
Phi-3.5 Vision	1.993	2.489	0.000	0.070	0.039
Llama-3.2 11B	2.988	3.807	0.001	0.081	0.051
Qwen2-VL 2B	1.394	1.599	0.013	0.075	0.063
GOT OCR base	0.956	1.174	0.011	0.064	0.047
Nayana-OCR	**0.270**	0.694	**0.248**	**0.740**	**0.516**
Marathi					
Tesseract	0.157	0.460	0.513	0.862	0.738
PaddleOCR	0.355	0.849	0.035	0.630	0.154
Phi-3.5 Vision	1.592	2.063	0.023	0.150	0.073
Llama-3.2 11B	2.421	2.724	0.007	0.108	0.074
Qwen2-VL 2B	1.251	1.269	0.069	0.248	0.181
GOT OCR base	0.915	0.988	0.021	0.095	0.060
Nayana-OCR	**0.143**	**0.457**	**0.540**	**0.866**	**0.753**
Gujarati					
Tesseract	**0.148**	0.446	**0.534**	**0.871**	**0.733**
PaddleOCR	0.800	0.914	0.026	0.124	0.068
Phi-3.5 Vision	3.329	3.008	0.006	0.091	0.047
Llama-3.2 11B	2.401	2.724	0.007	0.108	0.074
Qwen2-VL 2B	5.050	4.312	0.006	0.092	0.042
GOT OCR base	0.940	1.047	0.020	0.081	0.057
Nayana-OCR	0.172	**0.451**	0.476	0.839	0.707
Bengali					
Tesseract	0.241	0.590	0.259	0.738	0.492
PaddleOCR	0.704	0.798	0.014	0.096	0.029
Phi-3.5 Vision	2.041	2.110	0.008	0.014	0.042
Llama-3.2 11B	7.021	6.039	0.009	0.093	0.044
Qwen2-VL 2B	0.967	1.054	0.048	0.174	0.127
GOT OCR base	0.926	0.983	0.019	0.080	0.048
Nayana-OCR	**0.235**	**0.460**	**0.452**	**0.776**	**0.656**

A.2 Model Output Analysis

To evaluate the practical effectiveness of our model, we present a visual comparison between input documents and their corresponding Document Level OCR outputs. Figures 3 and 4 demonstrate the model's performance on Hindi and Bengali documents respectively.

The results demonstrate the model's robust performance across different Indic scripts. Note the preservation of both textual content and document structure in the generated outputs, highlighting the effectiveness of our approach in handling diverse document layouts and writing systems.


रूपांतरण न्यूट्रॉन के प्रसार से अपेक्षति GW - अजीब पदार्थ मध्यम वेगनेटकि क्षेत्रों की तुलना में दृढ़ता से असथरि है, और इसलए कुछ GW की उम्मीद है [21], एक अनशि्चति शक्त कि साथ। सुपरफ्लेयर के यूनसि मॉडल का ववरिण लुगोस एट अल [14] में चर्चा की गई है, जैसा कि पहले ही उल्लेख कयिा गया है, समग्र संभावनाएं सकारात्मक पता लगाने के लए उत्साहजनक लगती हैं, वही वचिार मोस्केरा क्यूस्टा एट अल [17] में प्रस्तुत मॉडल पर लागू होते हैं, भले ही अभी तक एसेरोशन-संचालति एसजीआर के लए सबूत नहीं मलि हैं [22]।

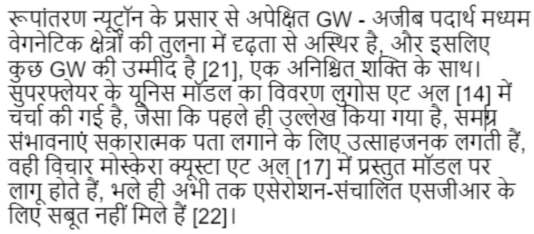
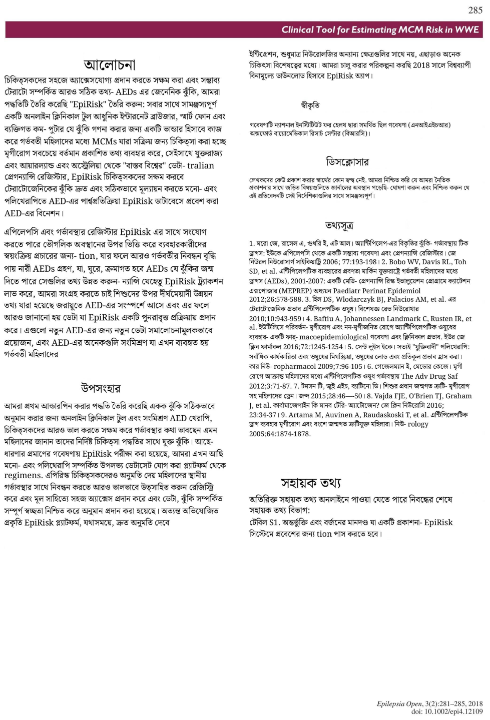

रूपांतरण न्यूटॉन के प्रसार से अपेक्षित GW - अजीब पदार्थ मध्यम वेगनेटिक क्षेत्रों की तुलना में दृढ़ता से अस्थिर है, और इसलिए कुछ GW की उम्मीद है [21], एक अनिश्चित शक्ति के साथ। सुपरफ्लेयर के यूनिस मॉडल का विवरण लुगोस एट अल [14] में चर्चा की गई है, जैसा कि पहले ही उल्लेख किया गया है, समग्र संभावनाएं सकारात्मक पता लगाने के लिए उत्साहजनक लगती हैं, वही विचार मोस्केरा क्यूस्टा एट अल [17] में प्रस्तुत मॉडल पर लागू होते हैं, भले ही अभी तक एसेरोशन-संचालित एसजीआर के लिए सबूत नहीं मिले हैं [22]।

Figure 3: Hindi Document Processing: Comparison between the original document (left) and the model's OCR output (right), demonstrating accurate text recognition and formatting preservation.


285

Clinical Tool for Estimating MCM Risk in WWE

Epilepsia Open, 3(2):281–285, 2018
doi: 10.1002/epi4.12109

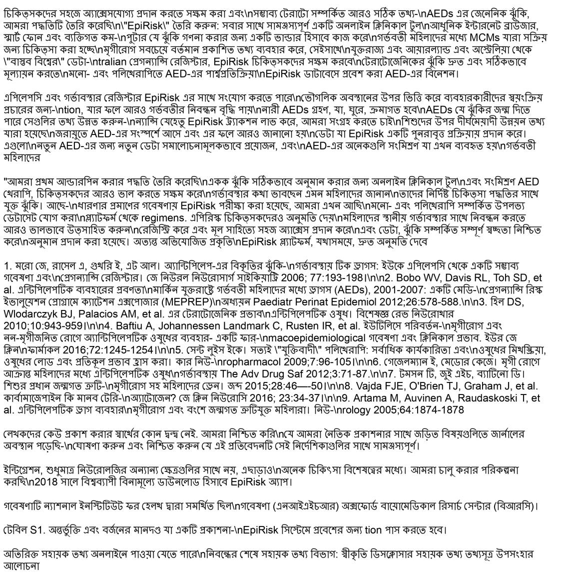
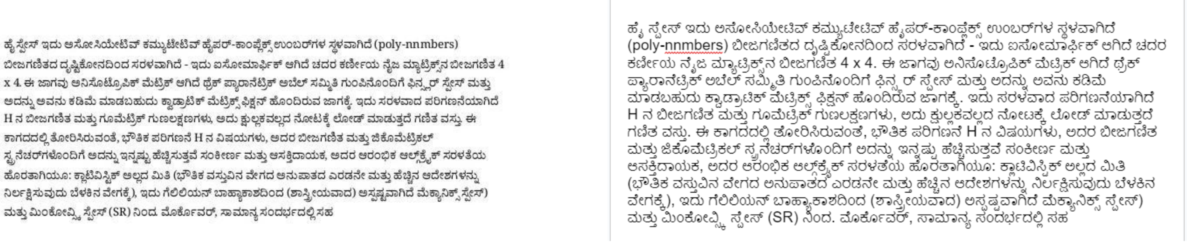


Figure 4: Bengali Document Processing: Visual comparison showing the model's capability to accurately process Bengali script while maintaining structural fidelity.


ಹೈ ಸ್ಪೇಸ್ ಇದು ಅಸೋಸಿಯೇಟಿವ್ ಕಮ್ಯುಟೇಟಿವ್ ಹೈಪರ್-ಕಾಂಪ್ಲೆಕ್ಸ್ ಉಂಬರ್‌ಗಳ ಸ್ಥಳವಾಗಿದೆ (poly-nnmbers)
ಬೀಜಗಣಿತದ ದೃಷ್ಟಿಕೋನದಿಂದ ಸರಳವಾಗಿದೆ - ಇದು ಐಸೋಮಾರ್ಫಿಕ್ ಆಗಿದೆ ಚದರ ಕರ್ಣೀಯ ನೈಜ ಮ್ಯಾಟ್ರಿಕ್ಸ್‌ನ ಬೀಜಗಣಿತ 4
x 4. ಈ ಜಾಗವು ಅನಿಸೊಟ್ರೊಪಿಕ್ ಮೆಟ್ರಿಕ್ ಆಗಿದೆ ಥ್ರೆಕ್ ಪ್ಯಾರಾನೆಟ್ರಿಕ್ ಅಬೆಲ್ ಸಮ್ಮಿತಿ ಗುಂಪಿನೊಂದಿಗೆ ಫಿನ್ಸ್ಲರ್ ಸ್ಪೇಸ್ ಮತ್ತು
ಅದನ್ನು ಅವನು ಕಡಿಮೆ ಮಾಡಬಹುದು ಕ್ವಾಡ್ರಾಟಿಕ್ ಮೆಟ್ರಿಕ್ಸ್ ಫಿಕ್ಷನ್ ಹೊಂದಿರುವ ಜಾಗಕ್ಕೆ. ಇದು ಸರಳವಾದ ಪರಿಗಣನೆಯಾಗಿದೆ
H ನ ಬೀಜಗಣಿತ ಮತ್ತು ಗೂಮೆಟ್ರಿಕ್ ಗುಣಲಕ್ಷಣಗಳು, ಅದು ಕ್ಷುಲ್ಲಕವಲ್ಲದ ನೋಟಕ್ಕೆ ಲೋಡ್ ಮಾಡುತ್ತದೆ ಗಣಿತ ವಸ್ತು. ಈ
ಕಾಗದದಲ್ಲಿ ತೋರಿಸಿರುವಂತೆ, ಭೌತಿಕ ಪರಿಗಣನೆ H ನ ವಿಷಯಗಳು, ಅದರ ಬೀಜಗಣಿತ ಮತ್ತು ಜಿಕೊಮೆಟ್ರಿಕಲ್
ಸ್ಟ್ರನೆಚರ್‌ಗಳೊಂದಿಗೆ ಅದನ್ನು ಇನ್ನಷ್ಟು ಹೆಚ್ಚಿಸುತ್ತವೆ ಸಂಕೀರ್ಣ ಮತ್ತು ಆಸಕ್ತಿದಾಯಕ, ಅದರ ಆರಂಭಿಕ ಆಲ್ಗ್ಕ್ರೈಕ್ ಸರಳತೆಯ
ಹೊರತಾಗಿಯೂ: ಕ್ಲಾಟಿವಿಸ್ಟಿಕ್ ಅಲ್ಲದ ಮಿತಿ (ಭೌತಿಕ ವಸ್ತುವಿನ ವೇಗದ ಅನುಪಾತದ ಎರಡನೇ ಮತ್ತು ಹೆಚ್ಚಿನ ಆದೇಶಗಳನ್ನು
ನಿರ್ಲಕ್ಷಿಸುವುದು ಬೆಳಕಿನ ವೇಗಕ್ಕೆ), ಇದು ಗೆಲಿಲಿಯನ್ ಬಾಹ್ಯಾಕಾಶದಿಂದ (ಶಾಸ್ತ್ರೀಯವಾದ) ಅಸ್ಪಷ್ಟವಾಗಿದೆ ಮೆಕ್ಯಾನಿಕ್ಸ್ ಸ್ಪೇಸ್)
ಮತ್ತು ಮಿಂಕೋವ್ಸ್ಕಿ ಸ್ಪೇಸ್ (SR) ನಿಂದ. ಮೊರ್ಕೋವರ್, ಸಾಮಾನ್ಯ ಸಂದರ್ಭದಲ್ಲಿ ಸಹ

ಹೈ ಸ್ಪೇಸ್ ಇದು ಅಸೋಸಿಯೇಟಿವ್ ಕಮ್ಯುಟೇಟಿವ್ ಹೈಪರ್-ಕಾಂಪ್ಲೆಕ್ಸ್ ಉಂಬರ್‌ಗಳ ಸ್ಥಳವಾಗಿದೆ
(poly-nnmbers) ಬೀಜಗಣಿತದ ದೃಷ್ಟಿಕೋನದಿಂದ ಸರಳವಾಗಿದೆ - ಇದು ಐಸೋಮಾರ್ಫಿಕ್ ಆಗಿದೆ ಚದರ
ಕರ್ಣೀಯ ನೈಜ ಮ್ಯಾಟ್ರಿಕ್ಸ್‌ನ ಬೀಜಗಣಿತ 4 x 4. ಈ ಜಾಗವು ಅನಿಸೊಟ್ರೊಪಿಕ್ ಮೆಟ್ರಿಕ್ ಆಗಿದೆ ಥ್ರೆಕ್
ಪ್ಯಾರಾನೆಟ್ರಿಕ್ ಅಬೆಲ್ ಸಮ್ಮಿತಿ ಗುಂಪಿನೊಂದಿಗೆ ಫಿನ್ಸ್ಲರ್ ಸ್ಪೇಸ್ ಮತ್ತು ಅದನ್ನು ಅವನು ಕಡಿಮೆ
ಮಾಡಬಹುದು ಕ್ವಾಡ್ರಾಟಿಕ್ ಮೆಟ್ರಿಕ್ಸ್ ಫಿಕ್ಷನ್ ಹೊಂದಿರುವ ಜಾಗಕ್ಕೆ. ಇದು ಸರಳವಾದ ಪರಿಗಣನೆಯಾಗಿದೆ
H ನ ಬೀಜಗಣಿತ ಮತ್ತು ಗೂಮೆಟ್ರಿಕ್ ಗುಣಲಕ್ಷಣಗಳು, ಅದು ಕ್ಷುಲ್ಲಕವಲ್ಲದ ನೋಟಕ್ಕೆ ಲೋಡ್ ಮಾಡುತ್ತದೆ
ಗಣಿತ ವಸ್ತು. ಈ ಕಾಗದದಲ್ಲಿ ತೋರಿಸಿರುವಂತೆ, ಭೌತಿಕ ಪರಿಗಣನೆ H ನ ವಿಷಯಗಳು, ಅದರ ಬೀಜಗಣಿತ
ಮತ್ತು ಜಿಕೊಮೆಟ್ರಿಕಲ್ ಸ್ಟ್ರನೆಚರ್‌ಗಳೊಂದಿಗೆ ಅದನ್ನು ಇನ್ನಷ್ಟು ಹೆಚ್ಚಿಸುತ್ತವೆ ಸಂಕೀರ್ಣ ಮತ್ತು
ಆಸಕ್ತಿದಾಯಕ, ಅದರ ಆರಂಭಿಕ ಆಲ್ಗ್ಕ್ರೈಕ್ ಸರಳತೆಯ ಹೊರತಾಗಿಯೂ: ಕ್ಲಾಟಿವಿಸ್ಟಿಕ್ ಅಲ್ಲದ ಮಿತಿ
(ಭೌತಿಕ ವಸ್ತುವಿನ ವೇಗದ ಅನುಪಾತದ ಎರಡನೇ ಮತ್ತು ಹೆಚ್ಚಿನ ಆದೇಶಗಳನ್ನು ನಿರ್ಲಕ್ಷಿಸುವುದು ಬೆಳಕಿನ
ವೇಗಕ್ಕೆ), ಇದು ಗೆಲಿಲಿಯನ್ ಬಾಹ್ಯಾಕಾಶದಿಂದ (ಶಾಸ್ತ್ರೀಯವಾದ) ಅಸ್ಪಷ್ಟವಾಗಿದೆ ಮೆಕ್ಯಾನಿಕ್ಸ್ ಸ್ಪೇಸ್)
ಮತ್ತು ಮಿಂಕೋವ್ಸ್ಕಿ ಸ್ಪೇಸ್ (SR) ನಿಂದ. ಮೊರ್ಕೋವರ್, ಸಾಮಾನ್ಯ ಸಂದರ್ಭದಲ್ಲಿ ಸಹ

Figure 5: Kannada Document Processing: Comparison between the original document (left) and the model's OCR output (right)

A.3 Training Dynamics Analysis

A.3.1 Single Language LoRA Adaptation

We first analyze the training dynamics for individual languages using LoRA with rank=64 and α=128. Figure 6 shows the training curves for Hindi and Kannada.

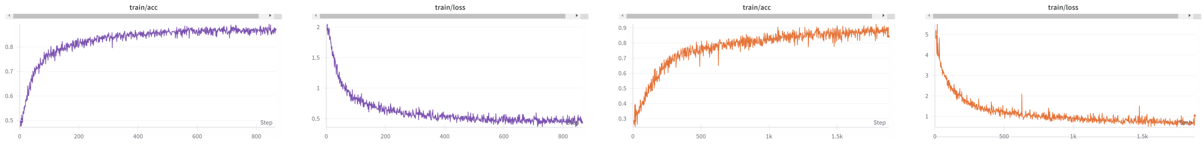


Figure 6: Single Language LoRA Training Dynamics: Training and loss curves for Hindi (purple) and Kannada (orange) using LoRA (r=64, α=128). Both languages show stable convergence patterns with Hindi achieving slightly faster convergence.

A.3.2 Multi-Language Joint Training

Building on the single language results, we investigate joint training on Hindi and Kannada. Figure 7 demonstrates the effectiveness of our multi-language approach.

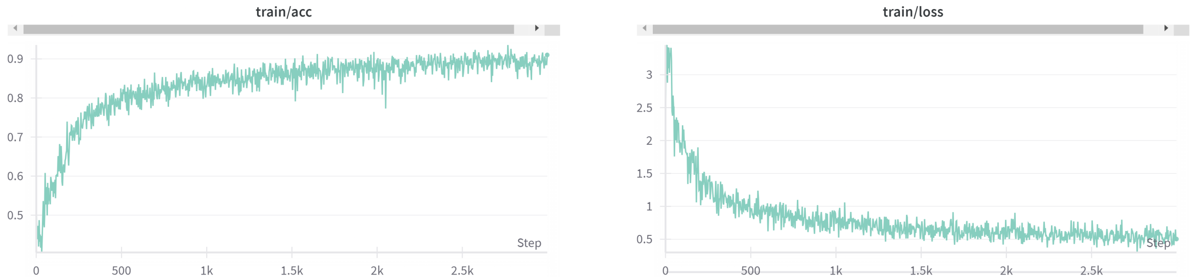


Figure 7: Joint Hindi-Kannada Training: The model maintains strong performance while learning both languages simultaneously, suggesting effective parameter sharing between related scripts.

A.3.3 Comparative Analysis of Joint vs Individual Training

To validate our multi-language approach, we compare joint training performance against individual language models. Figure 8 presents this critical comparison, where the orange line represents Kannada with rank 64 LoRA, the neon line shows the joint Hindi-Kannada LoRA (rank 64), and the green line indicates Hindi with rank 64 LoRA adaptation.

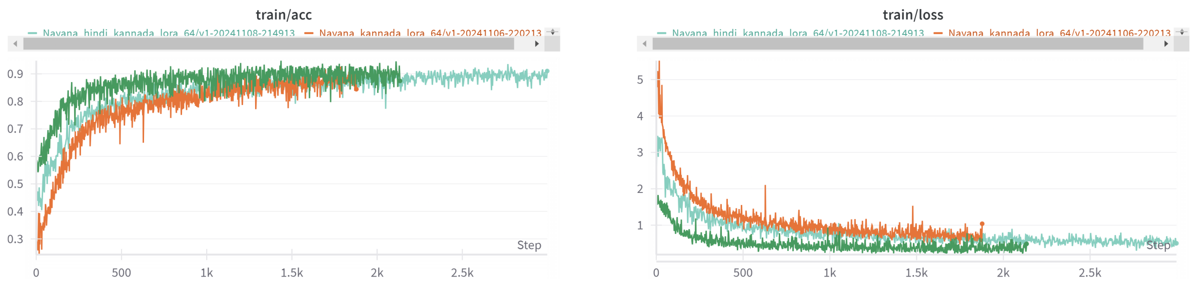


Figure 8: Comparative Analysis: Joint Hindi-Kannada training (r=64, α=128) versus individual language models. The joint model (neon) achieves comparable performance to individual Hindi (green) and Kannada (orange) models while using fewer parameters, demonstrating efficient cross-lingual transfer.

A.3.4 Vocabulary Expansion Experiments

Our initial experiments explored vocabulary expansion as a potential approach for handling multiple scripts. Figure 9 illustrates these challenges, comparing standard LoRA adaptation (purple lines) against vocabulary expansion attempts (grey lines).

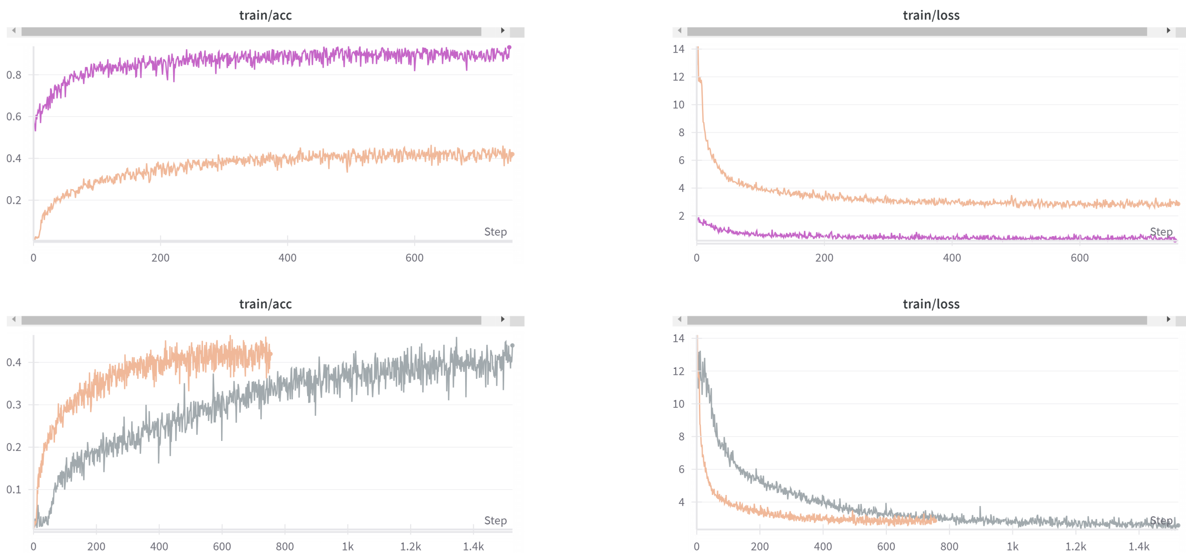


Figure 9: Vocabulary Expansion Analysis: Attempts to expand model vocabulary for Hindi showed poor convergence across different configurations. The standard vocabulary with LoRA adaptation (purple) proved more effective than expanded vocabulary approaches (grey), leading us to abandon the vocabulary expansion strategy.

A.4 Data Generation Examples

A.4.1 Page-Level Translation Examples

Our pipeline demonstrates robust translation capabilities while preserving document structure across all supported languages. Figures 10, 11, 12, 13 and 14 showcase these capabilities across different Indic scripts.

A.4.2 Section-Level Translation Examples

Figures 15, 16, 17 and 18 showcase section level translation capabilities across different Indic scripts.

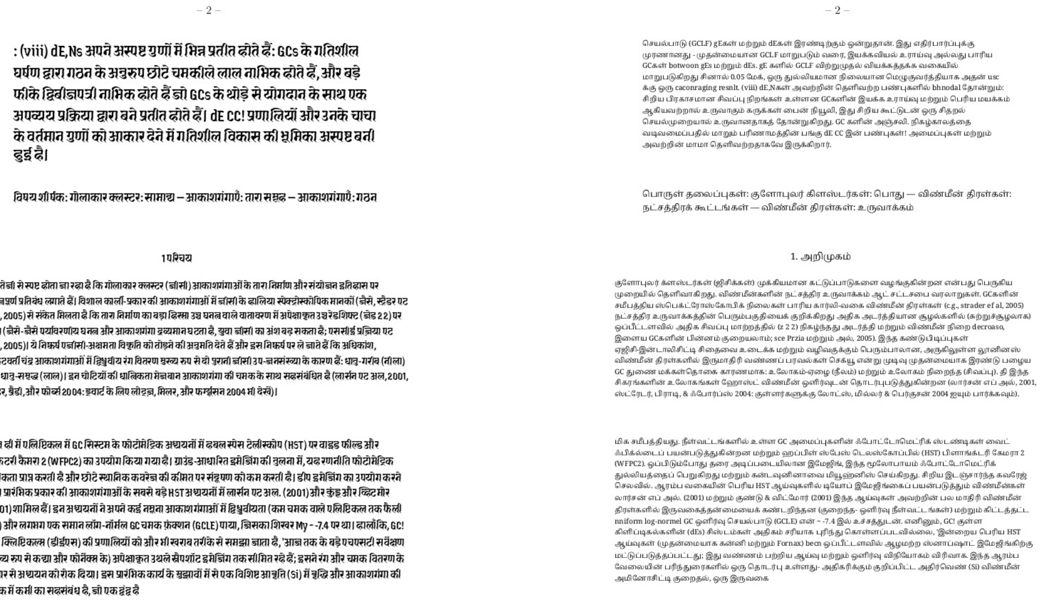


Figure 10: Page-level translation examples showing Hindi (left) and Tamil (right) translations with preserved document layout.

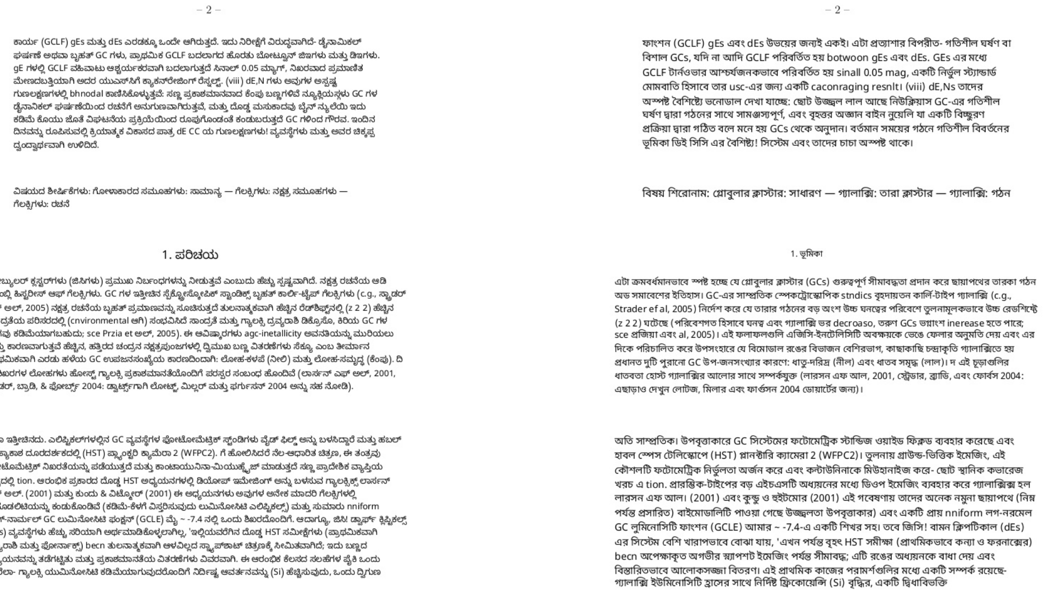


Figure 11: Page-level translation examples demonstrating Kannada (left) and Bengali (right) translations.

Figure 12: Page-level translation examples showing Malayalam (left) and Gujarati (right) translations.

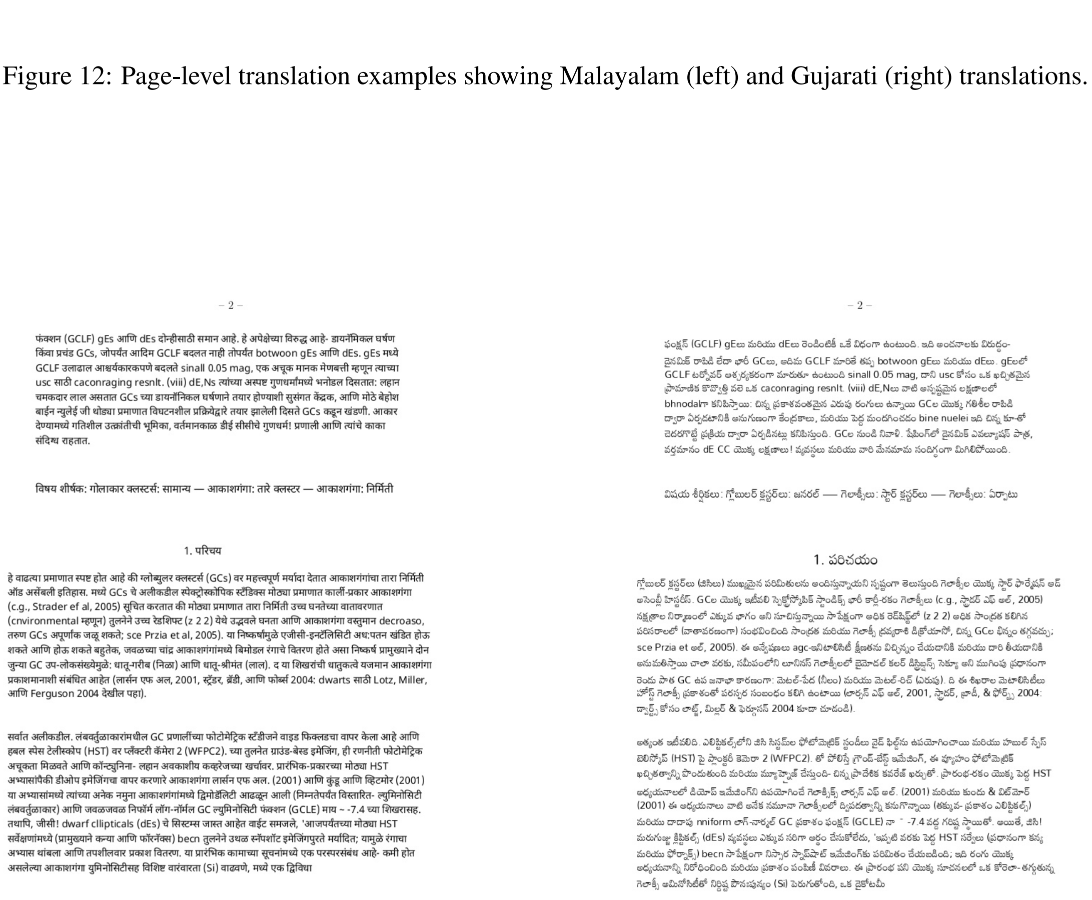
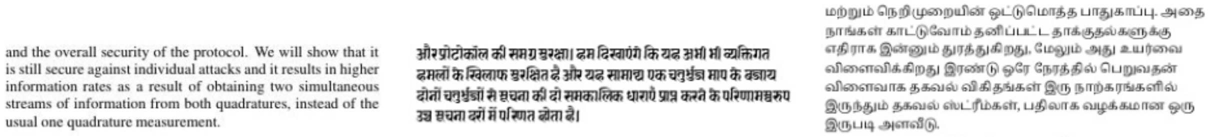


Figure 13: Page-level translation examples demonstrating Marathi (left) and Telugu (right) translations.

Figure 14: Page-level translation examples showing Odia (left) and Punjabi (right) translations.



and the overall security of the protocol. We will show that it is still secure against individual attacks and it results in higher information rates as a result of obtaining two simultaneous streams of information from both quadratures, instead of the usual one quadrature measurement.

Figure 15: Section-level translation examples showing English (original), Hindi, and Tamil translations.

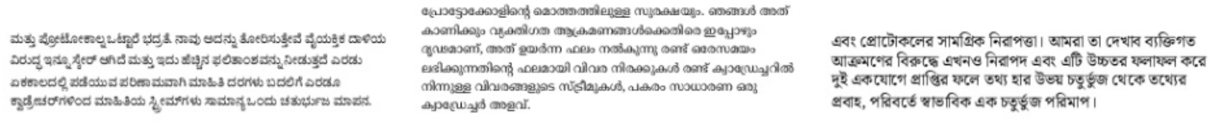


Figure 16: Section-level translation examples showing Kannada, Malayalam, and Bengali translations.

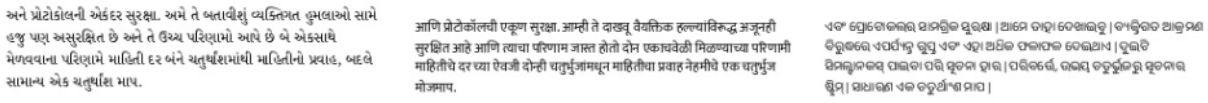


Figure 17: Section-level translation examples showing Gujarati, Marathi, and Odia translations.

Figure 18: Section-level translation examples showing Punjabi and Telugu translations.

A.5 Data Augmentation Examples

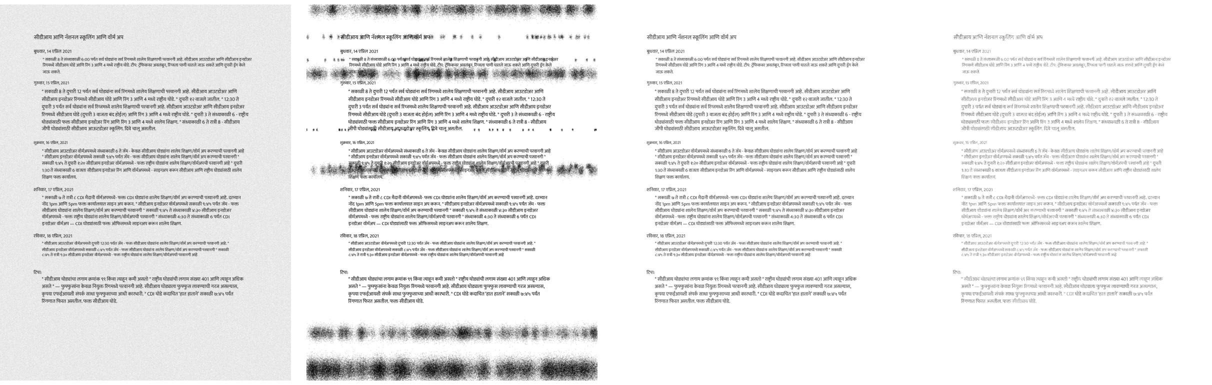


Figure 19: Document degradation examples showing (from left to right): background texturization, printer drum defects, ink mottling effect, and letterpress impression.

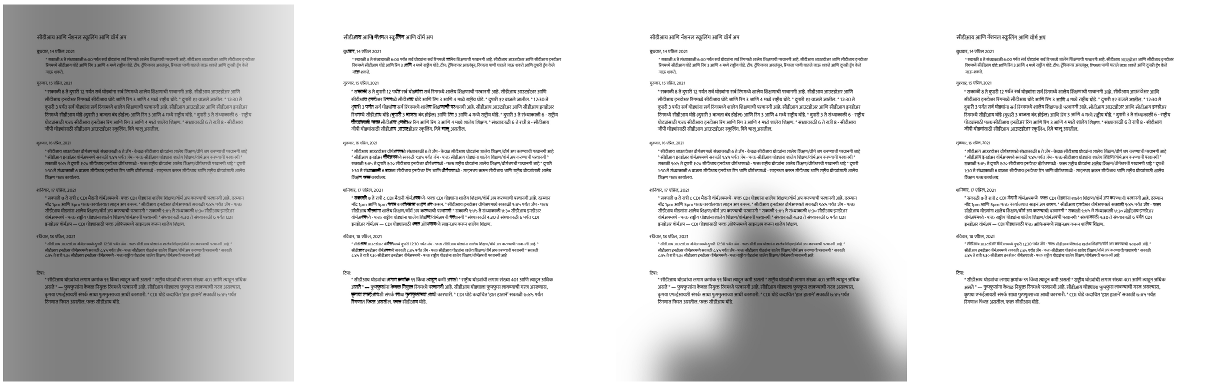


Figure 20: Document degradation examples showing (from left to right): lighting gradient, line degradation, shadow effects, and ink bleeding.