PROJECT REVIEW 1

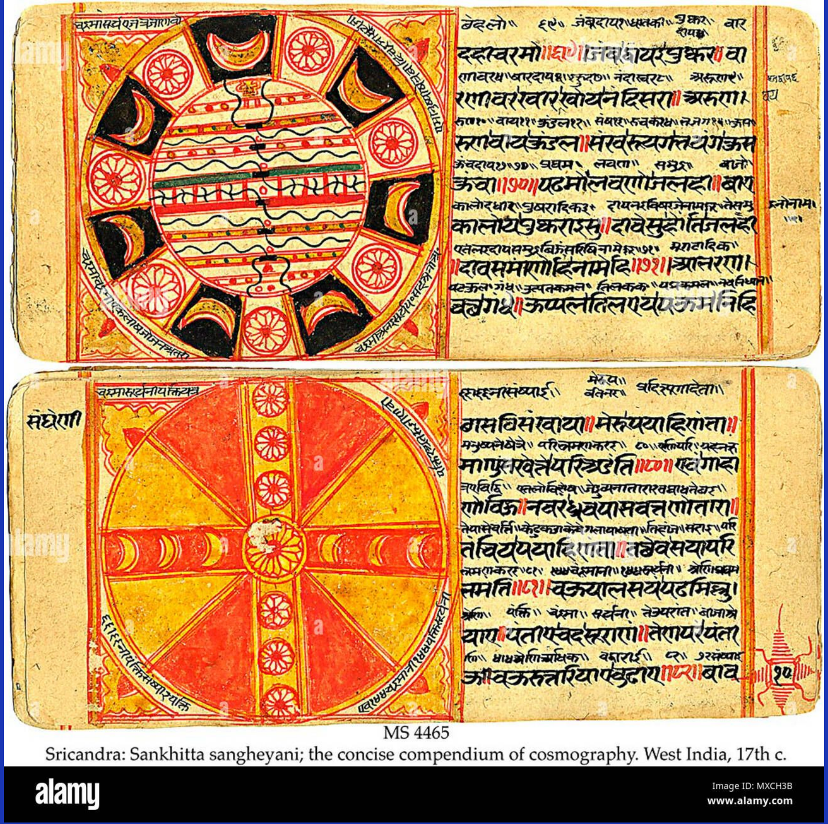
PraTranV3

Prakrit Languages Translator

Under the Guidance of

Mr. Kishor Pathak

Department of Information Technology, Vishwakarma Institute of Information Technology



Meet the Team

Sarvesh Chaudhari

PRN: 22210831
Roll No.: 332005

Sankalp Chakre

PRN: 22210615
Roll No.: 332004

Adwait Gondhalekar

PRN: 22211523
Roll no. : 332021

Nirdosh Chavhan

PRN: 22210030
Roll No.: 332011

Introduction

- Prakrit is an ancient Middle Indo-Aryan language.
- Used in Jain, Buddhist, and classical Indian texts.
- Maharashtri Prakrit (previous project) vs. Ardha Magadhi Prakrit (current project).
- **Why Ardha Magadhi?**
- Language of early Jain and Buddhist scriptures.
- No existing machine translation models.

Aim of the Project

Develop an Ardha Magadhi Prakrit to English machine translation system.

Fine-tune pre-trained LLMs (LLAMA, Claude, NLLB, M2M100, etc.) for accuracy.

Use quantization techniques (LoRA, QLoRA) for efficient model adaptation (since we are going to fine tune LLM).

Deploy a web application to make translation accessible to researchers and the public.

Current Challenges

- Extremely Low-Resource Language – No structured datasets.
- OCR Issues – Google Vision API misreads Devanagari Prakrit as English.
- Complex Grammar – Different from modern Indo-Aryan languages.
- Fine-tuning LLMs – Requires efficient training strategies due to high computational cost.

Objectives

- Bridge the gap in Prakrit-to-English machine translation.
- Improve OCR-based text extraction from historical documents.
- Optimize models using fine-tuning techniques (LoRA, QLoRA).
- Ensure accessibility via a web-based translation tool.

Literature Review

Title	Technology + Dataset	Methodology	Advantage	Limitation	Accuracy
From LLM to NMT: Advancing Low-Resource Machine Translation with Claude(2024)	LLMs (GPT models, Claude), Traditional NMT (MarianNMT), Dataset: FLORES-200, TED Talks	Few-shot prompting, RLHF	More fluent outputs	May miss domain-specific nuances	LLM BLEU: 28.0 vs. Baseline BLEU: 25.0
Neural Machine Translation for Low-Resource Languages from a Chinese-centric Perspective: A Survey(2024)	Multilingual Transfer Learning, Adapter Layers, Contrastive Learning, Dataset: CCMatrix, CWMT	Byte-Pair Encoding (BPE), Character-level representations	Effective for script-based low-resource languages	May not generalize to different scripts	BLEU: 30.0, TER: 5% reduction from baseline
Advances in Interactive Machine Translation with Large Language Models(2024)	Interactive LLMs (LLaMA, PaLM), Dataset: Synthetic datasets, UN Parallel Corpus	Human-in-the-loop, Dynamic feedback integration	Allows incremental improvement via user interaction	Less suited for batch processing	BLEU: 20.0 → 28.0 after feedback

8Literature Review

LLMs for Machine Translation in Medium-Resourced Languages: Transfer Abilities & Parallel Data Impact(2024)	Fine-tuning models (LLaMA-2, mBART), Dataset: OPUS, Multilingual Common Crawl	Cross-lingual transfer, Parallel data impact analysis	Effective for medium-resource languages	May not extend to very low-resource cases	LEU: 26, chrF++: 44, COMET: 0.68
NusaMT-7B: Machine Translation for Low-Resource Indonesian Languages with Large Language Models(2024)	7B Parameter LLM, LoRA for efficient tuning, Dataset: Indonesian Wikipedia, OSCAR Corpus	LoRA-based adaptation for efficiency	Computationally efficient, adaptable for other low-resource languages	Specific to Indonesian, may not generalize	BLEU: 31, chrF: 47
Segment-Based Interactive Machine Translation for Pre-trained Models(2024)	Pre-trained LLMs, Segment-based translation, Dataset: Europarl, TED Talks	Implements a feedback loop to refine translations iteratively	Reduces translation errors by 15-20%	Requires additional user feedback mechanisms	BLEU, Human Evaluation
Machine Translation Evaluation Metrics Benchmarking: From Traditional MT to LLMs(2023)	Evaluation frameworks, comparing traditional MT metrics (BLEU, METEOR) with LLM-based (COMET), Dataset: FLORES, WMT Test Sets	Mathematical analysis of scoring systems	Provides metric guidance for evaluating translation models	Does not focus on model development	COMET: 0.7, TER: 25, chrF++: 50

Literature Review

Neural Machine Translation for Low-Resource Languages: A Survey(2023)	Transformer-based NMT, Data Augmentation, Unsupervised Learning, Dataset: OPUS, WMT	Training paradigms: data augmentation, semi-supervised learning, zero-shot translation	Covers multiple NMT methods	Lacks implementation details for LLMs	BLEU varies across methods
A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models(2023)	LLMs (GPT-4, ChatGPT), Dataset: WMT, FLORES-200	Examines prompt-based translation vs. traditional NMT	Higher fluency and robustness	Challenges in domain-specific translations	Google (BLEU: 31.66), DeepL (BLEU: 31.22), Tencent (BLEU: 29.69), GPT-3.5 (BLEU: 24.73)
New Trends in Machine Translation using Large Language Models: Case Examples with ChatGPT(2023)	ChatGPT, Instruction-tuned models, Interactive MT, Dataset: Real-world MT datasets	Compares rule-based vs. data-driven approaches for MT	Highlights ChatGPT's interactive MT advantages	Inconsistencies in ChatGPT translations	BLEU varies based on prompt tuning
BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models(2023)	Transformer models, Instruction-tuned LLMs, Dataset: Multilingual translation datasets	Fine-tuning and interaction-based alignment	Improves cross-lingual alignment	Requires extensive fine-tuning	Outperforms baseline translation models by 3-5 BLEU points

Literature Review

Dict-NMT: Bilingual Dictionary-based NMT for Extremely Low-Resource Languages(2022)	NMT models + Bilingual Dictionaries, Dataset: Low-resource language pairs	Integrates bilingual dictionaries for translation improvement	Works for languages with no bilingual corpora	Needs quality bilingual dictionaries	BLEU-based evaluation (no exact score provided
Adapting the Tesseract OCR Engine for Tamil and Sinhala Legacy Fonts(2021)	Tesseract OCR, LSTM-based font recognition, Dataset: Tamil-Sinhala-English Parallel Corpus	Enhanced Tesseract for recognizing legacy fonts	Significant error reduction	Limited to printed text, no handwritten recognition	Error rate reduced (Tamil: 6.03% → 2.61%, Sinhala: 7.61% → 4.74%)
Sentence Level Alignment of Digitized Books Parallel Corpora	Alignment algorithms, Dataset: Digitized fiction & non-fiction books	Combines alignment algorithms with proactive learning	Improves precision in aligning translated texts	Effectiveness depends on text quality	Improved alignment precision (exact scores not provided)
Survey of Low-Resource Machine Translation(2018)	Statistical MT, NMT, Transfer Learning, Unsupervised MT, Dataset: OPUS, WMT	Multilingual modeling, adversarial training	Provides historical context for low-resource MT	Lacks recent advancements with LLMs	Transfer learning improves BLEU scores in low-resource pairs

Gap Analysis

1. **Focus on Extremely Low-Resource Languages:** Most studies overlook languages with minimal data, such as Ardha Magadhi, which lack parallel corpora for effective NMT or LLM adaptation.
2. **Domain-Specific Translation Issues:** While LLMs offer fluency, they often fail to handle specialized vocabulary, making them less reliable for niche or historical languages.
3. **OCR and Parallel Corpus Gaps:** There is limited research on integrating OCR tools and legacy font recognition for low-resource languages, which could enhance translation tasks.
4. **Interactive MT for Low-Resource Languages:** User-feedback-based improvements are mainly explored for widely spoken languages, with limited focus on how such methods can aid translations for extremely low-resource languages.

Feasibility

Data Availability

Collaboration with Bhandarkar Institute.

Technical Feasibility

Fine-tuning LLMs with quantization.

Computational Cost

Requires efficient LoRA-based tuning.

Deployment Feasibility

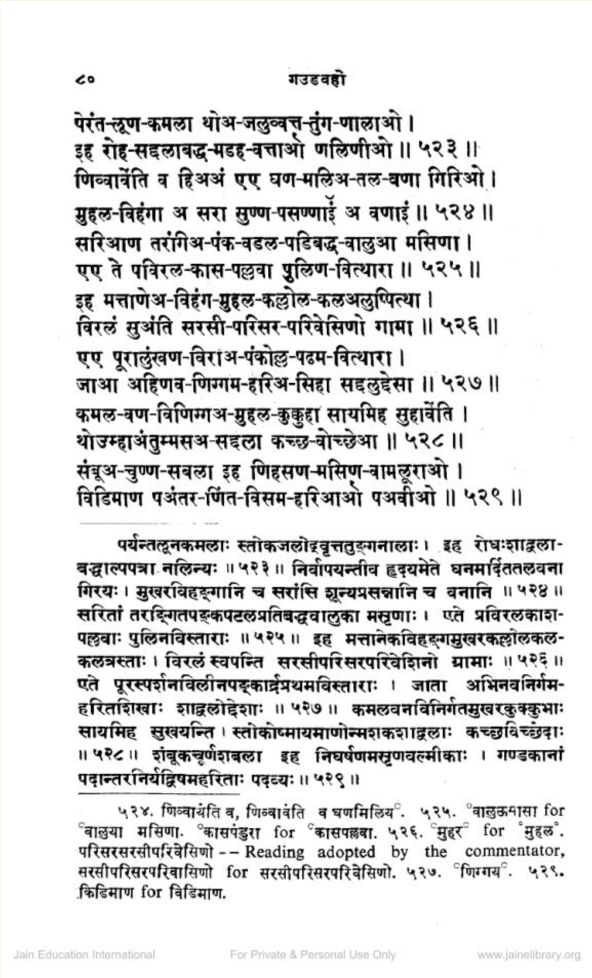
Web application ensures accessibility.

Dataset Creation

- Data Source: Bhandarkar Institute's ancient texts.
- OCR Extraction: Using Google Vision API to extract text from scanned documents.
- Challenges:
  - OCR misinterpretation of Devanagari text (converted into English incorrectly).
  - Manual intervention needed for text correction.

Dataset Issues

- OCR Issues:
- Example Issue:
    - <English text> (Devanagari Prakrit) <English text>
    - Google Vision API incorrectly converts Devanagari to English.
- Solution:
    - Preprocessing pipeline with manual correction & custom OCR models.



८० गउडवहो

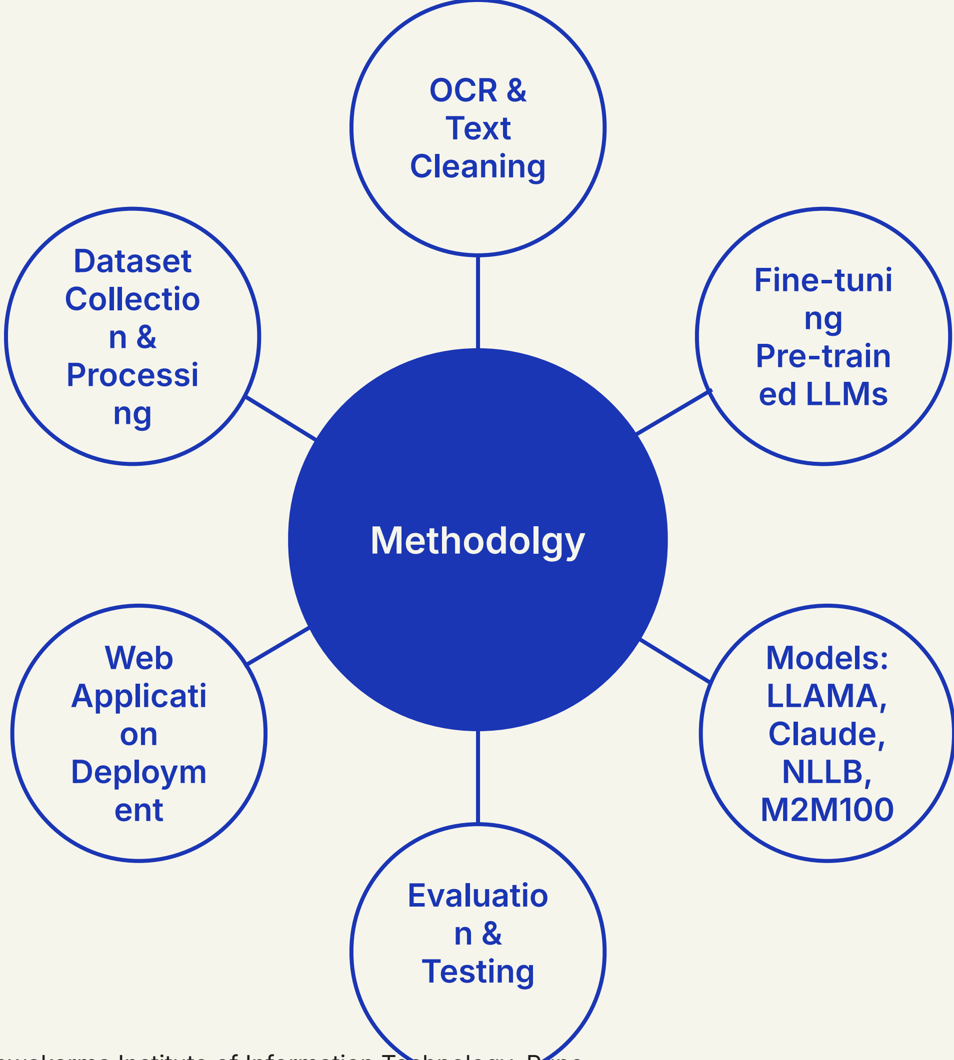
पेरंत-लूण-कमला थोअ-जलुव्वत्त-तुंग-णालाओ।
इह रोह-सद्दलाबद्ध-मडह-वत्ताओ णलिणीओ॥ ५२३॥
णिव्वावेंति व हिअअं एए घण-मलिअ-तल-वणा गिरिओ।
मुहल-विहंगा अ सरा सुण्ण-पसण्णाइँ अ वणाइं॥ ५२४॥
सरिआण तरंगिअ-पंक-वडल-पडिबद्ध-वालुआ मसिणा।
एए ते पविरल-कास-पल्लवा पुलिण-वित्थारा॥ ५२५॥
इह मत्ताणेअ-विहंग-मुहल-कल्लोल-कलअलुप्पित्था।
विरलं सुअंति सरसी-परिसर-परिवेसिणो गामा॥ ५२६॥
एए पूरालुंखण-विराअ-पंकोल्ल-पढम-वित्थारा।
जाआ अहिणव-णिग्गम-हरिअ-सिहा सद्दलुद्देसा॥ ५२७॥
कमल-वण-विणिग्गअ-मुहल-कुक्कुहा सायमिह सुहावेंति।
थोउम्हाअंतुम्मसअ-सद्दला कच्छ-वोच्छेआ॥ ५२८॥
संबूअ-चुण्ण-सबला इह णिहसण-मसिण-वामलूराओ।
विडिमाण पअंतर-णिंत-विसम-हरिआओ पअवीओ॥ ५२९॥

पर्यन्तलूनकमलाः स्तोकजलोद्वृत्ततुङ्गनालाः। इह रोधःशाद्वलाबद्धाल्पपत्रा नलिन्यः॥ ५२३॥ निर्वापयन्तीव हृदयमेते घनमर्दिततलवना गिरयः। मुखरविहङ्गानि च सरांसि शून्यप्रसन्नानि च वनानि॥ ५२४॥ सरितां तरङ्गितपङ्कपटलप्रतिबद्धवालुका मसृणाः। एते प्रविरलकाशपल्लवाः पुलिनविस्ताराः॥ ५२५॥ इह मत्तानेकविहङ्गमुखरकल्लोलकलकलत्रस्ताः। विरलं स्वपन्ति सरसीपरिसरपरिवेशिनो ग्रामाः॥ ५२६॥ एते पूरस्पर्शनविलीनपङ्कार्द्रप्रथमविस्ताराः। जाता अभिनवनिर्गमहरितशिखाः शाद्वलोद्देशाः॥ ५२७॥ कमलवनविनिर्गतमुखरकुक्कुभाः सायमिह सुखयन्ति। स्तोकोष्मायमाणोन्मशकशाद्वलाः कच्छविच्छेदाः॥ ५२८॥ शंबूकचूर्णशबला इह निघर्षणमसृणवल्मीकाः। गण्डकानां पदान्तरनिर्यद्विषमहरिताः पदव्यः॥ ५२९॥

५२४. णिव्वायंति व, णिव्वावंति व घणमिलिय°. ५२५. °वालुऊमासा for °वालुया मसिणा. °कासपंडुरा for °कासपल्लवा. ५२६. °मुहर° for °मुहल°. परिसरसरसीपरिवेसिणो -- Reading adopted by the commentator, सरसीपरिसरपरिवासिणो for सरसीपरिसरपरिवेसिणो. ५२७. °णिग्गय°. ५२९. किडिमाण for विडिमाण.

Methodology

- We are going to first extract text from the Text sources that have been collected
- From these extracted texts we have to create a parallel corpora.
- Using this Parallel Corpora we will fine tune LLMs or SMT using Quantization techniques like LoRA and QLoRA to save computation costs.
- Finally we will deploy the model on cloud and create an inference pipeline



Model Fine Tuning Strategy

- Base Models: LLAMA, Claude, M2M100, NLLB.
- Fine-Tuning Approach:
  - LoRA & QLoRA for low-memory fine-tuning.
  - Hyperparameter tuning (batch size, learning rate, epochs).
  - RLHF (Reinforcement Learning from Human Feedback)
  - Contrastive Learning
- Evaluation Metrics: BLEU, METEOR, chrF++, BERT score, COMET.

Thank You