

# Pathshala.ai: An Offline Bangla Language Model for High School Education in Bangladesh

**A Thesis for B.Sc. in Software Engineering, IICT, SUST**

Presented by-

Abraar Masud Nafiz and  
Hamim Rahman

We wouldn't be here today without the support, guidance and supervision of **Dr Ahsan Habib**, Associate Professor, IICT, SUST. So, We'd like to begin with a Thank you note - to our supervisor, faculties and board members of this presentation.

# Research background

- Educational disparity in Bangladesh: urban vs rural
- Increasing accessibility of smartphones and cheap devices
- Need for Bangla-specific models

- **Problem Description:**

- Lack of resources in rural areas
- High computational cost of existing NLP models
- Limited Bangla-oriented tools

# Objectives

- Develop a Bangla NLP model aligned with the NCTB curriculum.
- Optimize for low-resource devices without internet connectivity.
- Benchmark performance and usability.
- Bridge the education gap with an accessible AI-driven tutor.

# Research Significance

- Contributes to **SDG 4: Quality Education**
- Addresses linguistic and technological challenges
- Provides a scalable solution for rural education
- Sets a framework for similar low-resource models globally

# Technical Foundations

## **Core Concepts -**

- Transformer-based architectures (e.g., BERT, GPT)
- Challenges specific to Bangla (morphology, syntax, data scarcity)
- Importance of lightweight NLP models.

## **Optimization Techniques -**

- Knowledge distillation
- Quantization
- Sparse attention mechanisms

# Before we begin, our first problem was - Data!

## **Data Sources -**

- NCTB textbooks for classes 6–10
- Educational platforms (Shikkhok Batayon with 11k publications, BanglaWiki)
- Practice materials and past exam papers.

## **Preprocessing Steps -**

- Tokenization, normalization, and lemmatization
- Removal of irrelevant content
- Annotation for curriculum alignment

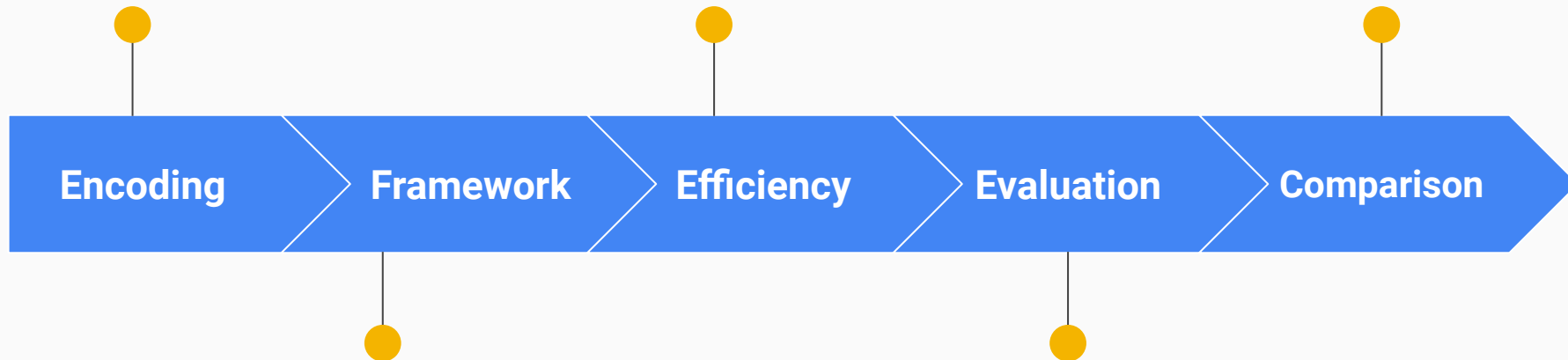
**Now, An Overview Of  
Pathshala.ai**



**Morphology-aware tokenization,  
Context-sensitive,  
Syntax-conscious positional  
encoding**

**Quantization and  
Knowledge Distillation**

**With BanglaBERT, mBERT,  
DistilBERT, and GPT-2 Bangla.  
Our model has the fast  
inference time and smaller size  
while sacrificing quality.**



- Transformer-based with 12 encoder layers
- Sparse attention for computational efficiency

**Performance Metrics -**

- Perplexity: 14.2, Accuracy: 70%, BLEU-4: 45.3%

**Efficiency Metrics -**

- Model size: 120 MB, Inference time: 1.1–10s

# Before and After Custom Tokenization

হরকা মারার হাসি-আজি উড়ে তরা  
শিকার ক্ষণতি অসীমান,  
স্পমায় শসতের করিত জ্বানচি।  
উঠেছে তারে কোথাও তাসেরে প্রাণপান।  
তারার অঙ্গুল হল নূরহী হাসি  
আপন নাই মধুলা।  
শেষে  
যা ছড়াই, ' চায় গেছে বিচিত্রা হল কিরুণ—  
' শেষবরণীপ  
সামার রকমা পাগল পায়,  
উদ্ নের টানায় রে পোড়া  
ভাকে শিয়ে

Step	Description	Output
1	Input Sentence	বাংলাদেশে শিক্ষার মানোন্নয়নে প্রযুক্তির ব্যবহার ক্রমবর্ধমান গুরুত্বপূর্ণ হয়ে উঠছে
2	BPE Regex Pre-tokenization	Tokens are split based on BPE regex rules
3	Tokens	['বাংলাদেশে', 'শিক্ষার', 'মানোন্নয়নে', 'প্রযুক্তির', 'ব্যবহার', 'ক্রমবর্ধমান', 'গুরুত্বপূর্ণ', 'হয়ে', 'উঠছে']
4	Check Morph Table?	Decision to check if morph decomposition is needed
5a	Morph Decomposition	'মানোন্নয়নে' → ['মান', 'উন্নয়ন', 'নে'] 'গুরুত্বপূর্ণ' → ['গুরু', 'ত্ব', 'পূর্ণ']
5b	BPE Tokenizer	Tokenizes using standard BPE rules
6	Output Tokens	['বাংলা', 'দেশে', 'শিক্ষা', 'মান', 'উন্নয়ন', 'নে', 'প্রযুক্তি', 'ব্যবহার', 'ক্রম', 'বর্ধ', 'মান', 'গুরু', 'ত্ব', 'পূর্ণ', 'হয়ে', 'উঠছে']
7	Final Tokenization Output	Processed tokens ready for downstream tasks

## Example Prompts

```
inputs = tokenizer([
    a_prompt.format(
        "সিন্ধুসভ্যতা কোন সময়ের? ", # instruction
        "", # input
        "", # output
    )
], return_tensors = "pt").to("cuda")
```

```
outputs = model.generate(**inputs, max_new_tokens = 512, use_cache = True)
tokenizer.batch_decode(outputs)
```

Instruction: সিন্ধুসভ্যতা কোন সময়ের?

### Input:

### Response:

সিন্ধু উপত্যকা ১৩০০ থেকে ৩৩০০ খ্রীস্টাব্দে ভারত উপমহাদেশে খ্রীস্টাব্দে।

বর্তমান পাকিস্তান, আফগানিস্তান এবং বাংলাদেশের অংশগুলির মধ্যে অবস্থিত।

```
inputs = tokenizer([
    a_prompt.format(
        "ভিটামিন A এর অভাবে কোন সমস্যা দেখা দেয়?", # instruction
        "", # input
        "", # output
    )
], return_tensors = "pt").to("cuda")
```

```
outputs = model.generate(**inputs, max_new_tokens = 512, use_cache = True)
tokenizer.batch_decode(outputs)
```

Instruction: ভিটামিন A অভাবে কোন সমস্যা দেখা দেয়?

Input:

Response:

এর অভাব একটি গুরুতর স্বাস্থ্য সমস্যা হতে যা বিভিন্ন এবং মানসিক সমস্যা সৃষ্টি করতে পারে। ভিটামিন এ এর যুক্ত কিছু সাধারণ মধ্যে রয়েছে:

১. হ্রাস: ভিটামিন এ রাতের দৃষ্টিশক্তি বজায় রাখতে এবং চোখের রাখতে গুরুত্বপূর্ণ। এ এর অভাব রাতের চোখের ক্ষতি দিকে পরিচালিত করতে পারে।
২. ত্বকের সমস্যা: এ ত্বকের রাখতে সহায়তা এবং এটির অভাব ত্বকের যেমন ত্বক, ত্বকের রোগ এবং ত্বকের ক্যান্সারের বাড়িয়ে তুলতে পারে।
৩. শিশুদের বিকাশের ভিটামিন এ গুরুত্বপূর্ণ, বিশেষত তাদের দৃষ্টিশক্তি, হাড় এবং ত্বকের বিকাশের জন্য। এ এর বিলম্ব, দৃষ্টিশক্তি হ্রাস এবং হাড়ের সমস্যা সৃষ্টি করতে পারে।
৪. ভিটামিন এ এর অভাব রক্তের স্বাস্থ্যের জন্য গুরুত্বপূর্ণ।

রক্তের স্বাস্থ্যের জন্য গুরুত্বপূর্ণ।

এটি রক্তের স্বাস্থ্যের জন্য গুরুত্বপূর্ণ।

রক্তের স্বাস্থ্য

# Performance Overview

**Perplexity:** Training: 12.5 | Validation: 14.2

*Good generalization; reasonable confidence in predictions.*

**BLEU-1: 78.4%, BLEU-4: 45.3%**

*Relevance and coherence in generated text.*

**ROUGE-1: 80.1%, ROUGE-L: 75.3%**

*Covers key points comprehensively.*

**Accuracy:** 70% overall; Bangla: 92%, Science: 82%, Math: 30%.

**Computational Efficiency - Inference Time:** 1.1–10s | **Peak RAM:** 240 MB | **Model Size:** 150 MB, **CPU Utilization:** 75% during inference.

## Comparative Analysis

- **Pathshala.ai vs. Other Models:**
  - Smaller size (120 MB) and faster (2–10s) with 70% accuracy.
  - Comparisons: BanglaBERT (88%, 5.5s), mBERT (82%, 6.0s), GPT-2 Bangla (78%, 7.0s).

# Challenges and Limitations

- Poor performance in mathematical reasoning
- Scarcity of training data for some subjects
- Multi-modal Interaction (Which will be our first priority in future)

# Future Work

- **Expand dataset coverage**
- **Enhance numerical reasoning capabilities**
- **Incorporate audio-based features for the visually impaired.**
- **Scale deployment across rural areas**
- **Further optimize model efficiency with advanced techniques**

# Conclusion

## **Achievements -**

- Developed a lightweight, curriculum-aligned Bangla language model
- Focusing on rural education with minimal resources, sacrificing everything over computational efficiency.

## **Impact -**

- Bridges educational gaps
- Offers a scalable framework for low-resource NLP applications

## **Future Vision -**

- Transform education for millions of rural students in Bangladesh and beyond.

Thank You!