



# A Comprehensive Approach for English(Bangla phonetic)-to-Bangla Transliteration Using Deep Learning

---

## Team

**Md Mizbah Uddin Junaed**

**Marajul Islam Shawn**

Supervised by,

**Mohammad Abdullah Al Mumin, PhD**

Professor,

Department of CSE, SUST

# Outline

**01 Introduction**

**02 Why this study?**

**03 Related Works**

**04 Goals & Challenge**

**05 Applications**

**06 Our Approach**

**07 Datasets**

**08 Proposed Model**

**09 Evaluation Metrics**

**10 Results**

## What is transliteration?

- Phonetic conversion of words across a pair of languages  
(eg. English to Bengali)
- Unlike translation, it preserve sound/phonetics
- 2 types : Forward Transliteration and Backward Transliteration
- A simple example of bangla backward transliteration:

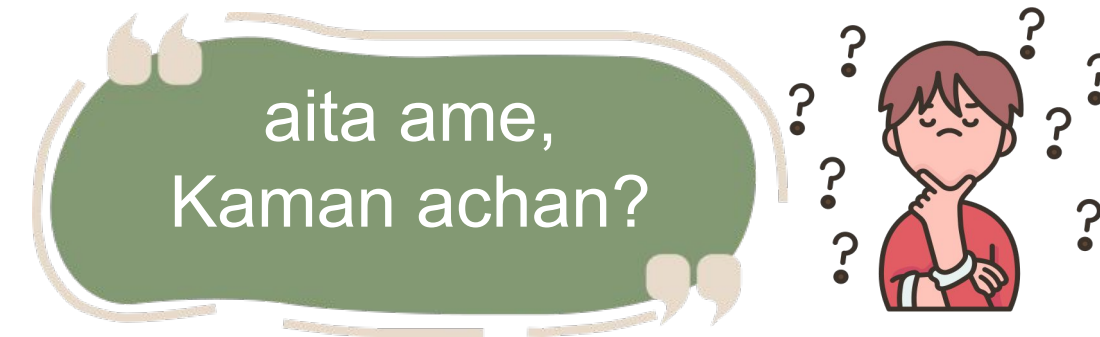
Ajker akash ta besh sundor

আজকের আকাশটা বেশ সুন্দর

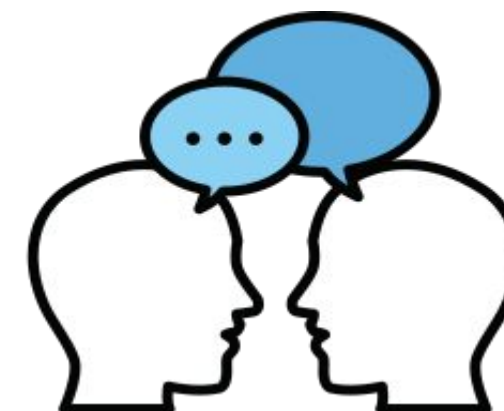
# Why this study?

Let's look at the scenarios :

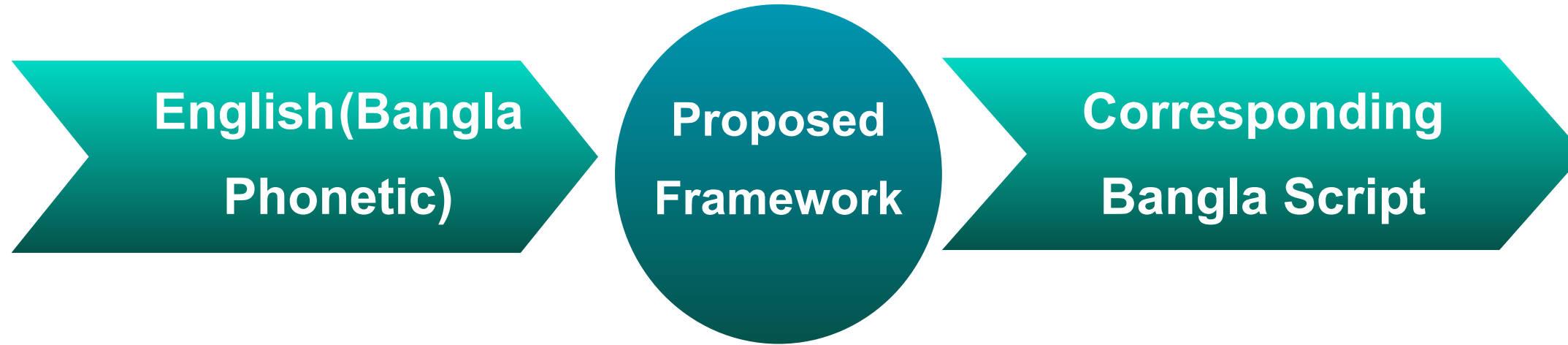
- Understanding native phonetic text (eg. **kemon, kaman, kamon**)
- OOV(Out-of-vocabulary) words problem
- You don't know other language well or, want to write something in bangla **but no input method!**
- To enhance cross-lingual communication
- May help in language learning



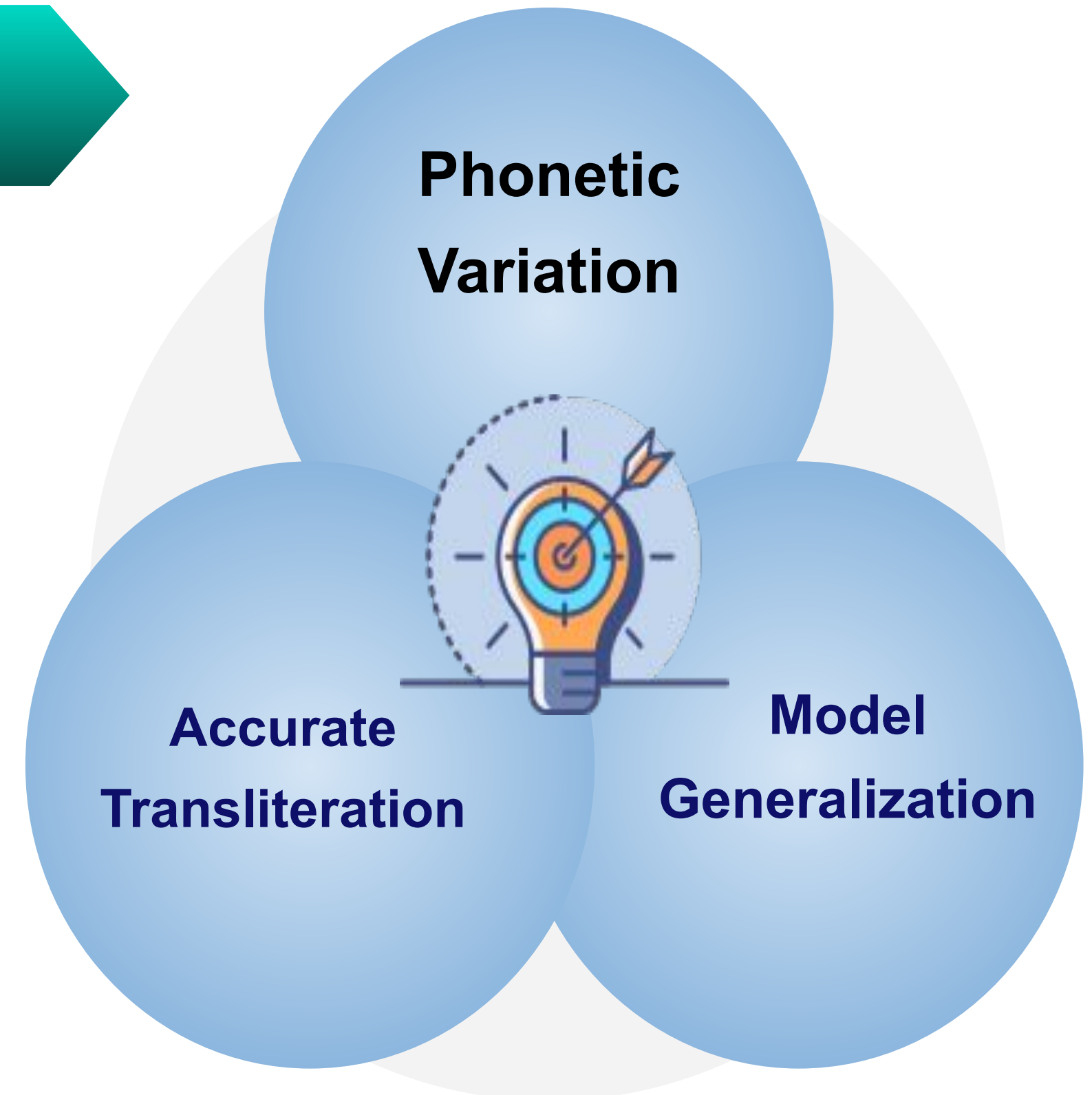
**No bangla layout  
available!**



# Goals and Challenges



- **More accurate bangla transliteration for diverse spelling**
- **Cross-Linguistic Adaptability**
- **User-Friendly Interaction & Continuous Improvement**

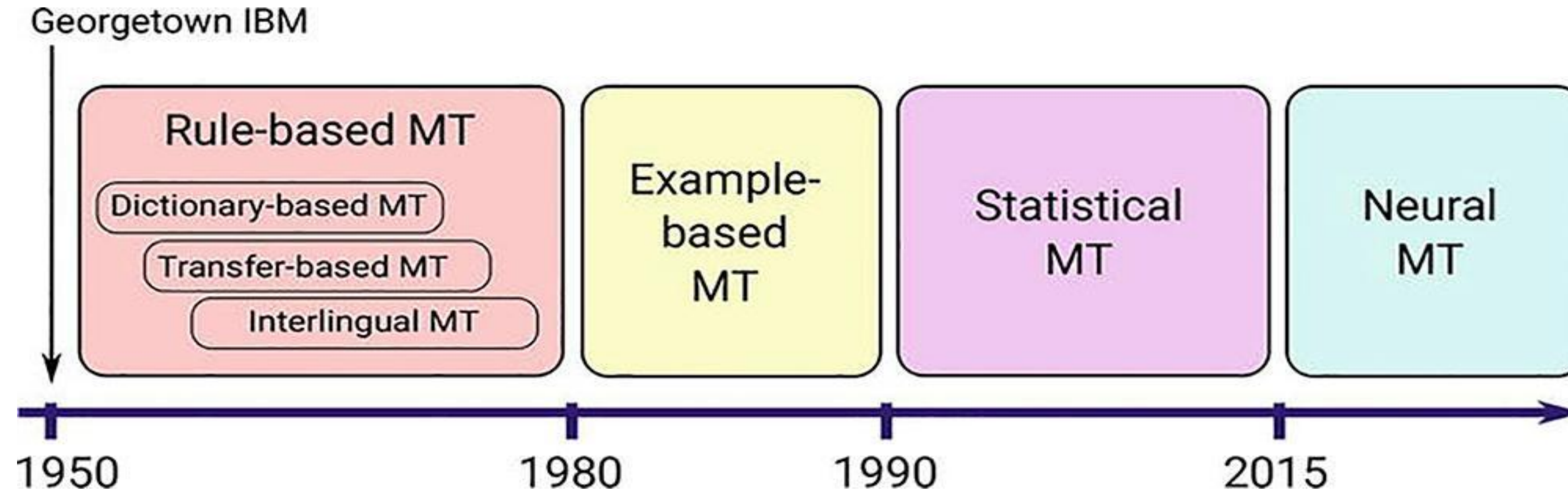


## Related Works

Studies Analysed about 29 related and partially related research works

- A research team from CSE, SUST worked on Bangla transliteration using a statistical approach in 2008
- Globally there are lots of works on **Hindi, Chinese, Arabic, French, Urdu and other languages**, no comprehensive works on bangla transliteration.
- Some research works on bengali with word based, phrase based etc.
- Most of the studies didn't handle all cases including **phonetic variations, native spelling etc.**

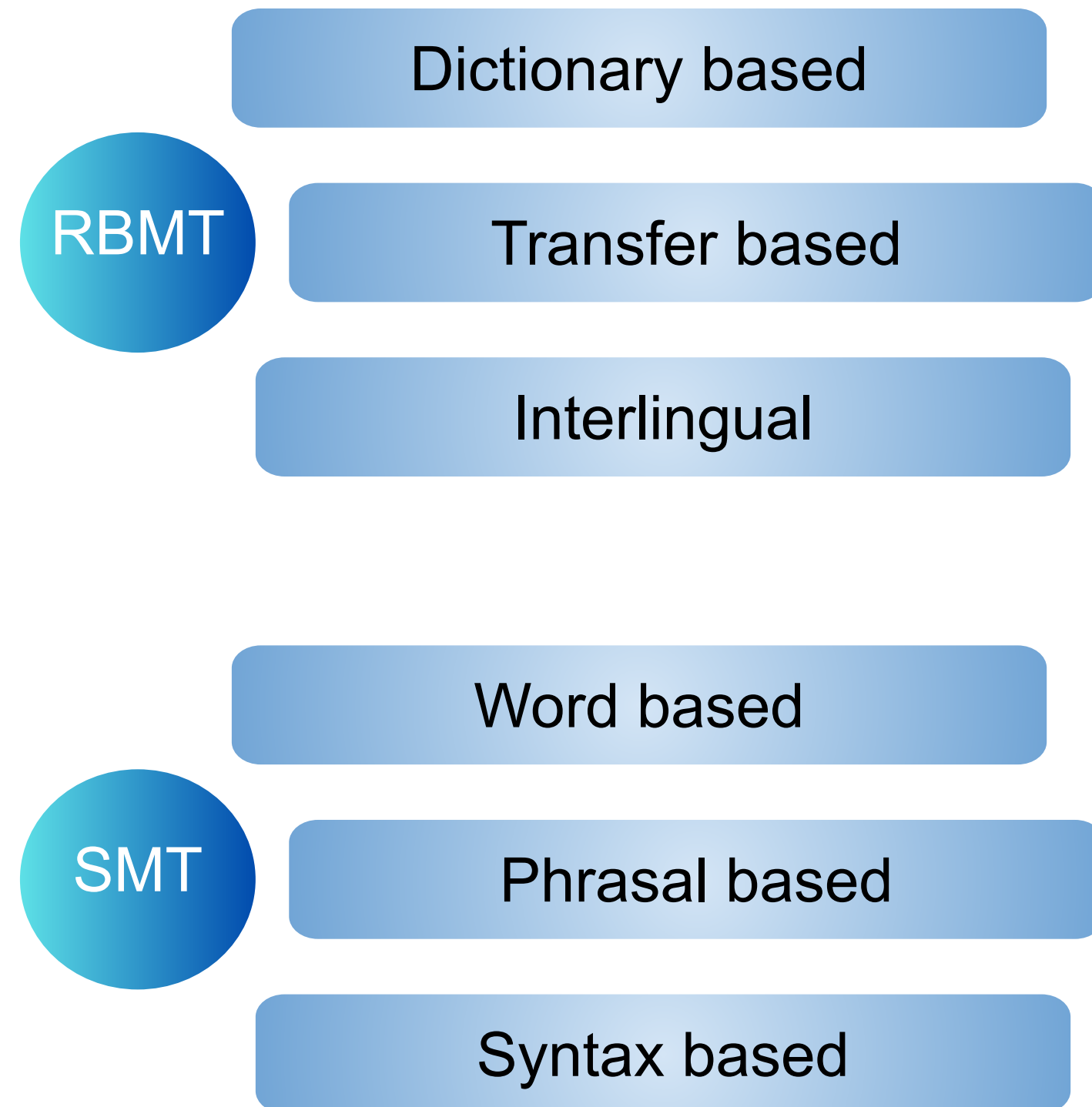
# Related Works



- Most of them are on **RMT** and **SMT**
- Some studies used NMT but did not handle all aspects.  
eg. **only works with name entities, phrase based, single word mapping etc.**



## Related Works



Gaps in the literature on RBMT based works

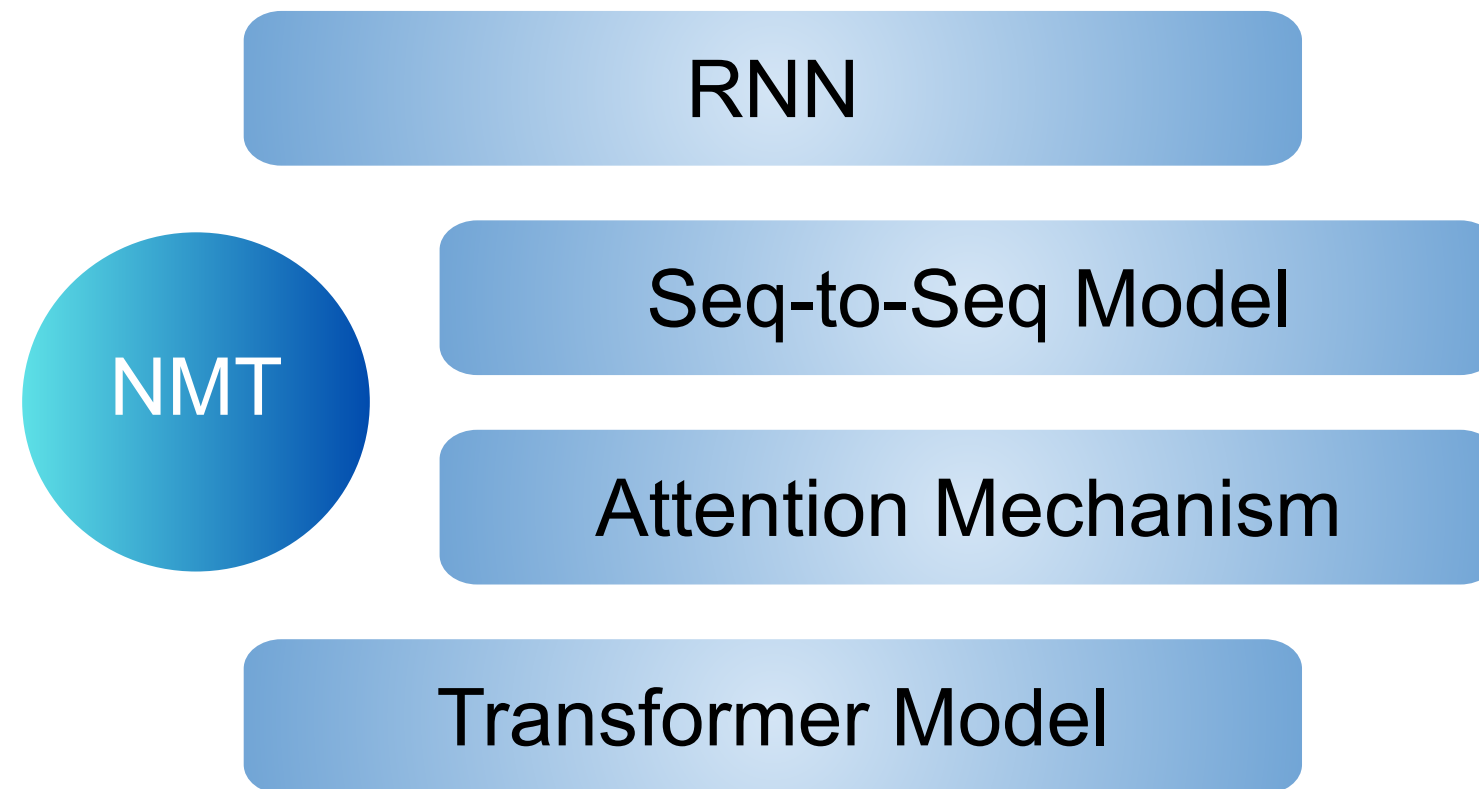
- One to one mapping
- Coverage and vocabulary
- Lack of generalization

Gaps in the literature on SMT based works

- Data dependency
- Context sensitivity
- Lack of learning
- Model complexity



## Related Works



Gaps in the literature of NMT based works

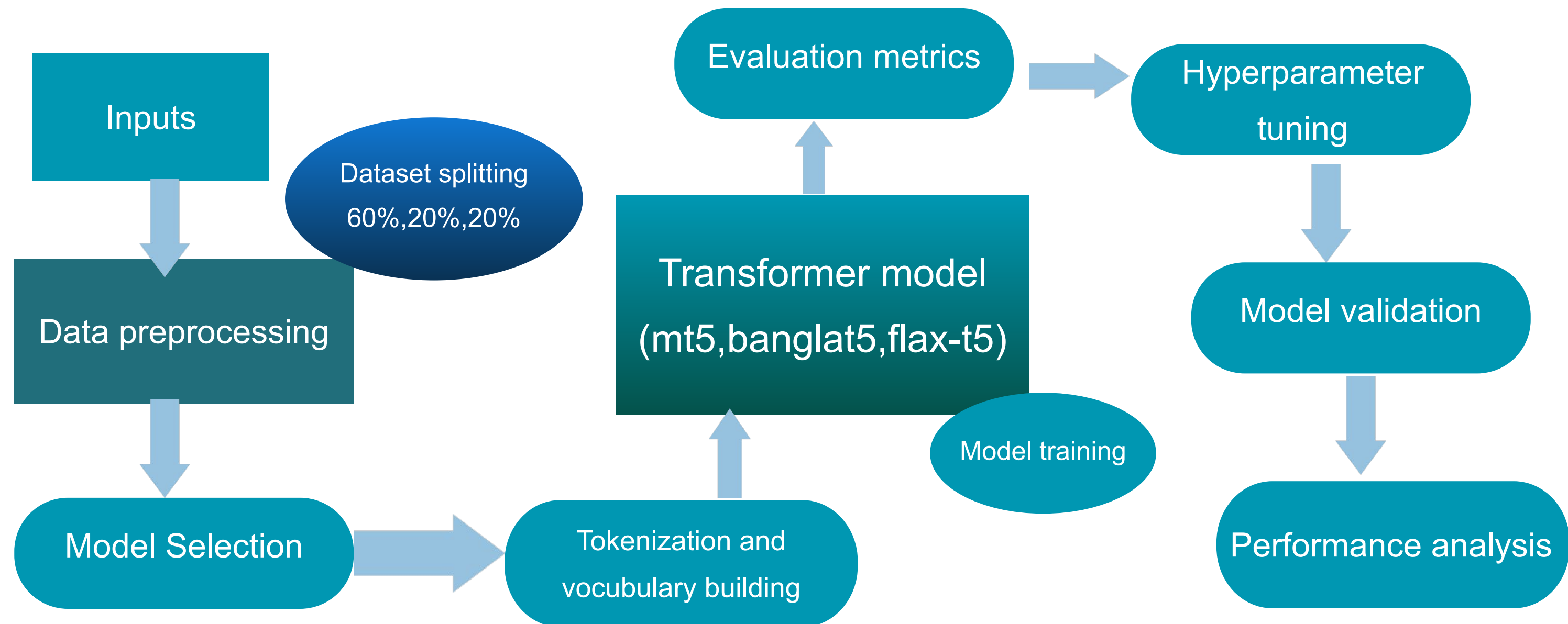
- Data Dependence
- Out-of-Vocabulary Challenges
- Linguistic Complexity

- Some tools that can be used for transliteration :

**Bengali Phonetic Parser, pyAvroPhonetic, Google Translate, BNTRANSLIT**

Drawback: Cannot handle native spelling & more nuanced cases effectively

## The Proposed Approach



# Dataset collection:

- We collected and prepared dataset consists of **9,000 sentence pairs**.
- Each pair has **Banglish (collected)** and **corresponding bangla script (human annotated)**
- **The dataset contains 45,000+ word pairs**
- **A robust dataset for better transliteration (From diverse sources including social media contents, chats, blogs etc.)**

File Edit View Insert Format Data Tools Extensions Help		
<div> <div>🔍 ↶ ↷ 🖨 📄 100%   £ % .0 .00 123   Defaul...   - 10 +   B I ⌂ A</div> </div>		
C13	fx	
	A	B
1	<b>English transliterated text (Bangla phonetic)</b>	<b>Corresponding Bangla Script</b>
2	shuvoratri ekti surjo abar osto jacche	শুভরাত্রি একটি সূর্য আবার অস্ত যাচ্ছে
3	Khusi houar jnno dhonnobad	খুশি হওয়ার জন্য ধন্যবাদ
4	mamara ghum theke ki ke uthla?	মামারা ঘুম থেকে কি কেউ উঠলো?
5	sob jhamela amar sathei hoi ken	সব ঝামেলা আমার সাথেই হয় ক্যান
6	bochorer ei anondomoi dinti tar sobtuku ronger tumai rangiye dik	বছরের এই আনন্দময় দিনটি তার সবটুকু রঙে তোমায় রাঙিয়ে দিক
7	pa rekhe kemon jeno nijer shohor nijer shohor mone hocche	পা রেখে কেমন যেন নিজের শহর নিজের শহর মনে হচ্ছে
8	ektu sokal bela	একটু সকাল বেলা
9	ekti raat er jonmo hocche	একটি রাত এর জন্ম হচ্ছে
10	utsober protti muhurto hoye udhuk aro ujjol. eid mubarak	উৎসবের প্রতিটি মুহূর্ত হয়ে উঠুক আরো উজ্জ্বল। ঈদ মুবারক
11	shudhui khobor, tar mohaproyan	শুধুই খবর, তাঁর মহাপ্রয়াণ
12	dinta sobar katok valo..	দিনটা সবার কাটুক ভালো ..
13	futlo abar vurer alo,,	ফুটলো আবার ভোরের আলো,,
14	sobai duya korben	সবাই দোয়া করবেন

# Proposed Model

**Transformer  
models**

**T5 (text-to-text  
transformer model)**

**banglat5**

**Flax\_counmmity t5**

**mt5-base**

**Developed by Buet Cse NLP  
group which is specially  
fine-tuned and trained for  
Bengali language**

**Flax-t5 is a variant of t5 model  
which has also encoder and  
decoder layers ; developed by  
google**

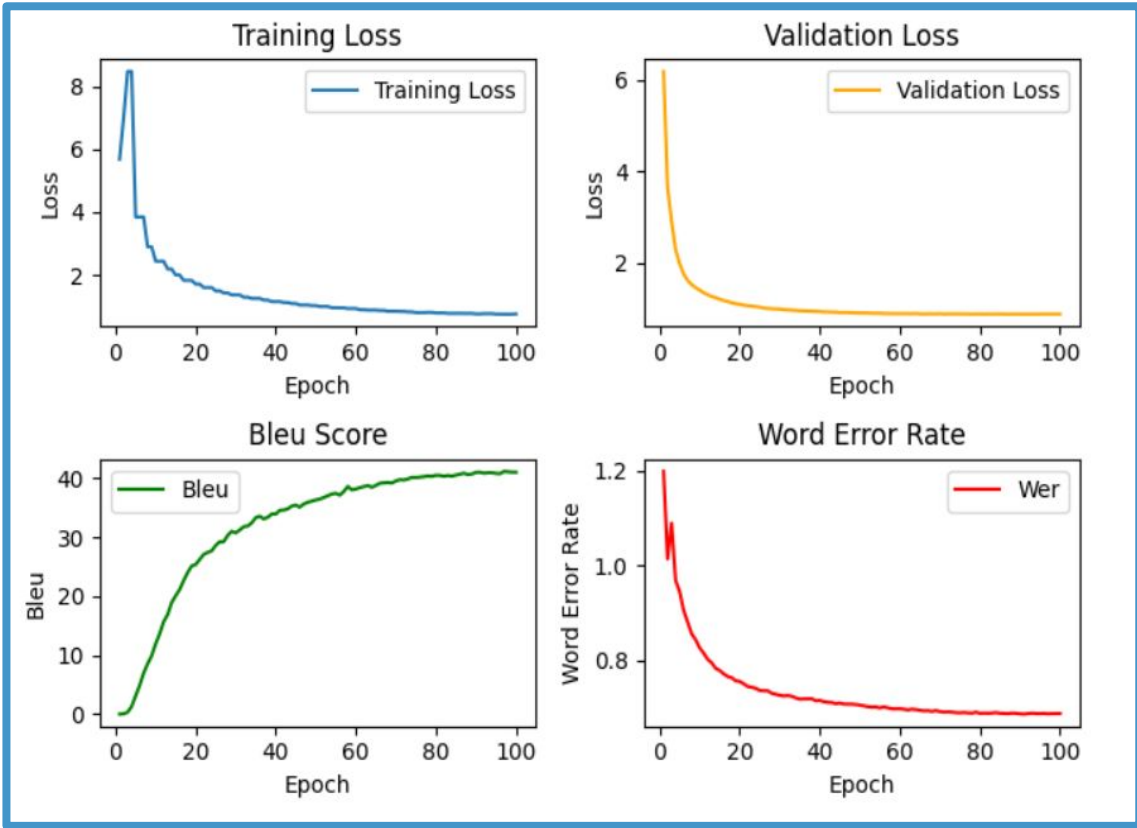
**Multilingual Translation  
Transformer model developed by  
google which capture long range  
dependence**

# Training session performance :

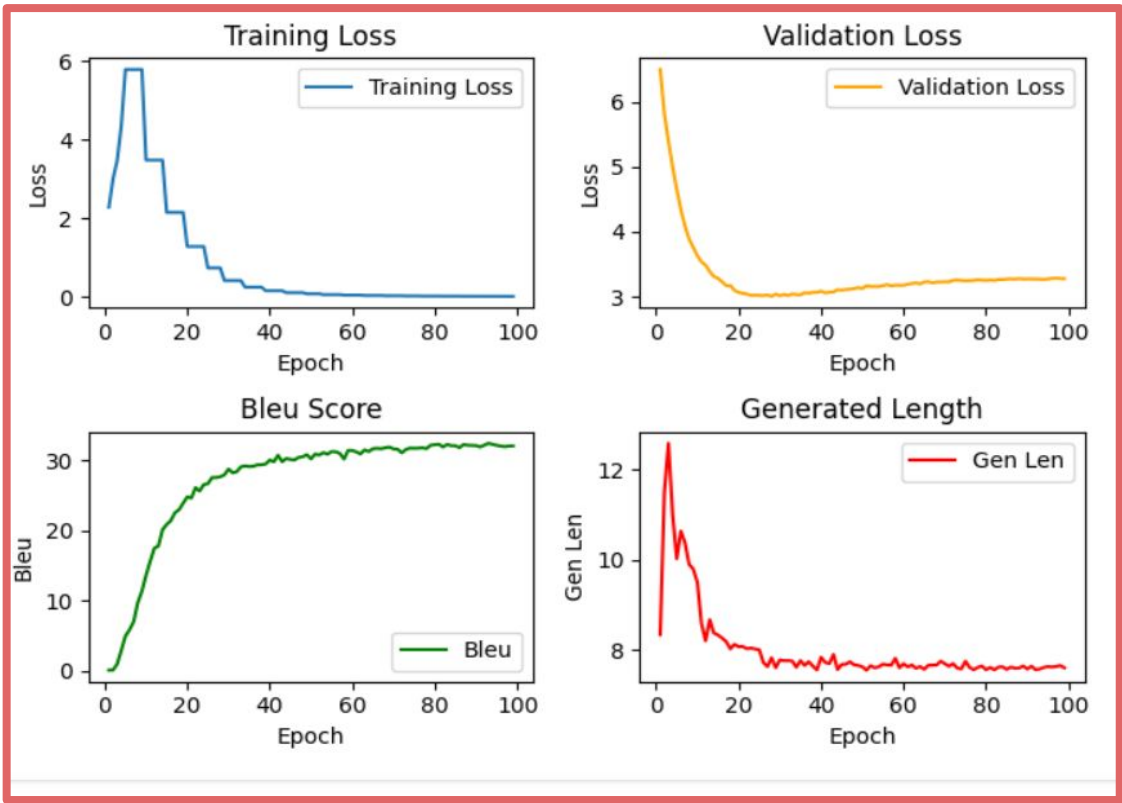
# Evaluation

**BLEU Score** : Bilingual Evaluation Understudy Similarity between machine translation and reference translation.

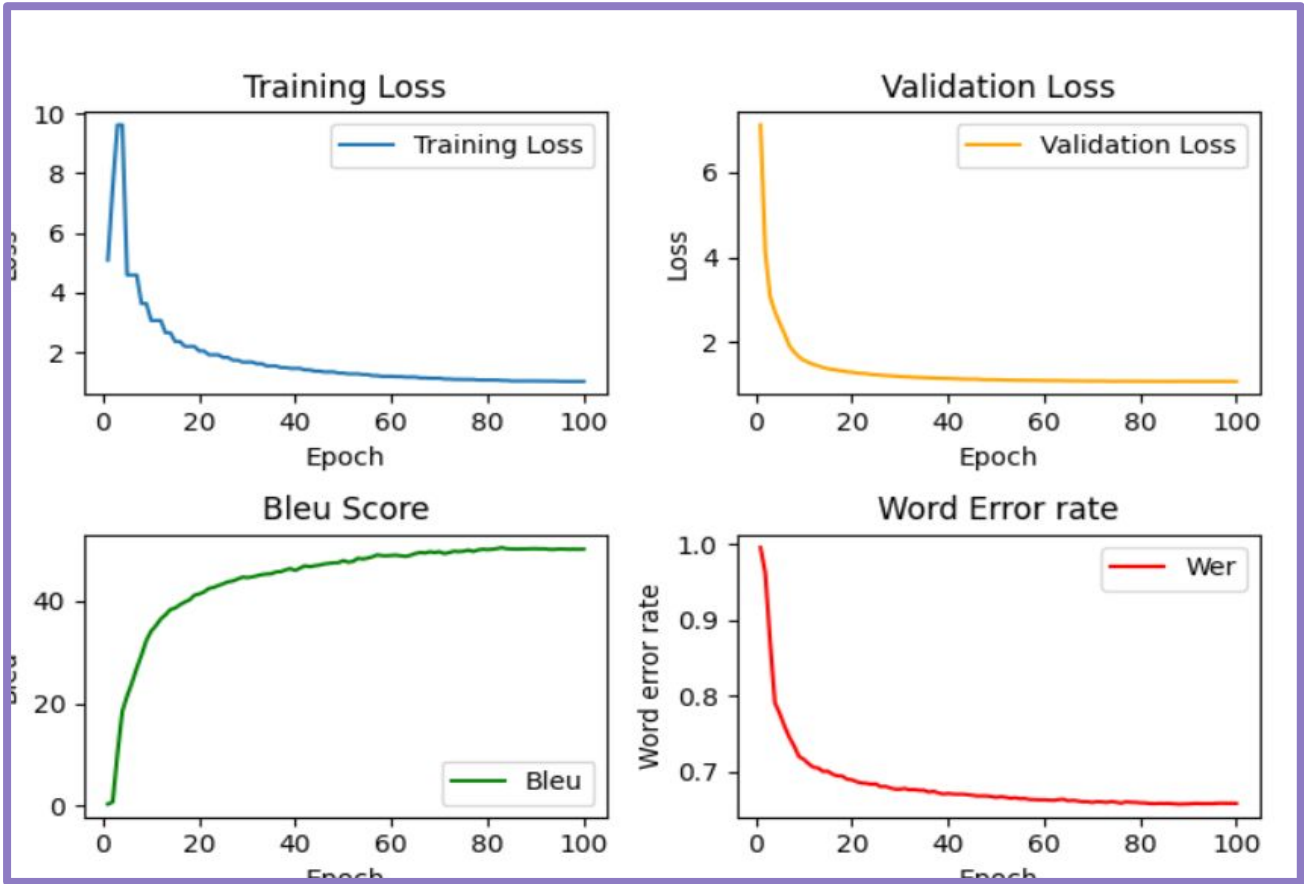
**Word Error Rate** : Metric that quantifies the difference between translated text and reference text by measuring the proportion of incorrectly recognized words



mt5



flax-t5



banglat5



Here is an output module:

Enter a text (input)	Bengali text (output)
kalke amar test porikkhar result prokash hobe.	কাল আমাদের টেস্ট পরিষ্কার রেজাল্ট প্রকাশ করা হবে।
khujji tomay shopnogaane. gudhuli belay ek cup cha, r artcell er amar pothchola	খুঁজি তোমায় স্বপ্নগান, গোধূলী বেলায় এক কাপ চা, আর আর্টসেল এর আমার পথচলা
nashokotar hortat, varate piketar diye bumabuzi, tobe prodhanmontrir sense of humor valo, shikar kortei hobe, toronder idol andur vashon dekhon bepok moja paisi	নাশকতার হরতাল, ভাড়া পিকিটার দিয়ে বোমাবাজি, তবে প্রধানমন্ত্রীর সেন্স অব হিউমার ভাল, শীকার করতেই হবে, তরুণদের আইডল আন্ডুর ভাষণটা দেখুন ব্যপুক মজা পাইছি

## Output & Performance

### Performance Analysis on testing

Model	Bleu Score	Word Error Rate
mt5	41.0128	0.6872
flax-t5	32.1055	0.7859
banglat5	46.0088	0.6580

# Applications

## Applications Cases

**Language  
Learning Aid**

**Improved Understanding  
and communication**

**Cultural Exchange**

**Dataset will be useful  
for future work**



# Thank You!



## Question and Answer...