

Natural Language Processing: Assignment 2

Team Members:

Nimitt

Pratham Sharda

Pranjal Gaur

Shubham Aggarwal

Harshit

Chirag Patel

GitHub Repository

Link to the GitHub repo containing all the files: https://github.com/Nimitt-nim/NLP_Assignment2

Tokenizer Training

Link to Tokenizers. : <https://www.kaggle.com/models/nimittnimizz/bengali-tokenizers>

Tokenizer	Fertility Score	Training Dataset Size (GB)	Tokenizer Type
1	1.36	0.54	Sentence BPE
2	1.10	0.62	Sentence BPE
3	1.33	0.40	Sentence BPE
4	1.09	1.56	Sentence BPE
5	1.45	1.10	BPE
6	1.16	0.89	BPE

Pretraining

3 LLAMA Models were trained using the best tokenizer. Out of them Model 1 performed the best and performance has been documented below:

Model 1

Link to Model: <https://www.kaggle.com/models/nimittnimizz/bengali-llama-45m>

Configuration

```
LlamaConfig {
  "_attn_implementation_autoset": true,
  "_name_or_path": "/Users/nimitt/Documents/NLP/NLP_Assignment2",
  "architectures": [
    "LlamaForCausalLM"
  ],
  "attention_bias": false,
  "attention_dropout": 0.0,
  "bos_token_id": 2,
  "eos_token_id": 3,
  "head_dim": 64,
  "hidden_act": "silu",
  "hidden_size": 512,
  "initializer_range": 0.02,
  "intermediate_size": 256,
  "max_position_embeddings": 2048,
  "mlp_bias": false,
  "model_type": "llama",
  "num_attention_heads": 8,
  "num_hidden_layers": 8,
  "num_key_value_heads": 8,
  "pad_token_id": 0,
  "pretraining_tp": 1,
  "rms_norm_eps": 1e-06,
  "rope_scaling": null,
```

```
"rope_theta": 10000.0,  
"tie_word_embeddings": false,  
"torch_dtype": "float32",  
"transformers_version": "4.46.2",  
"use_cache": true,  
"vocab_size": 32769  
}
```

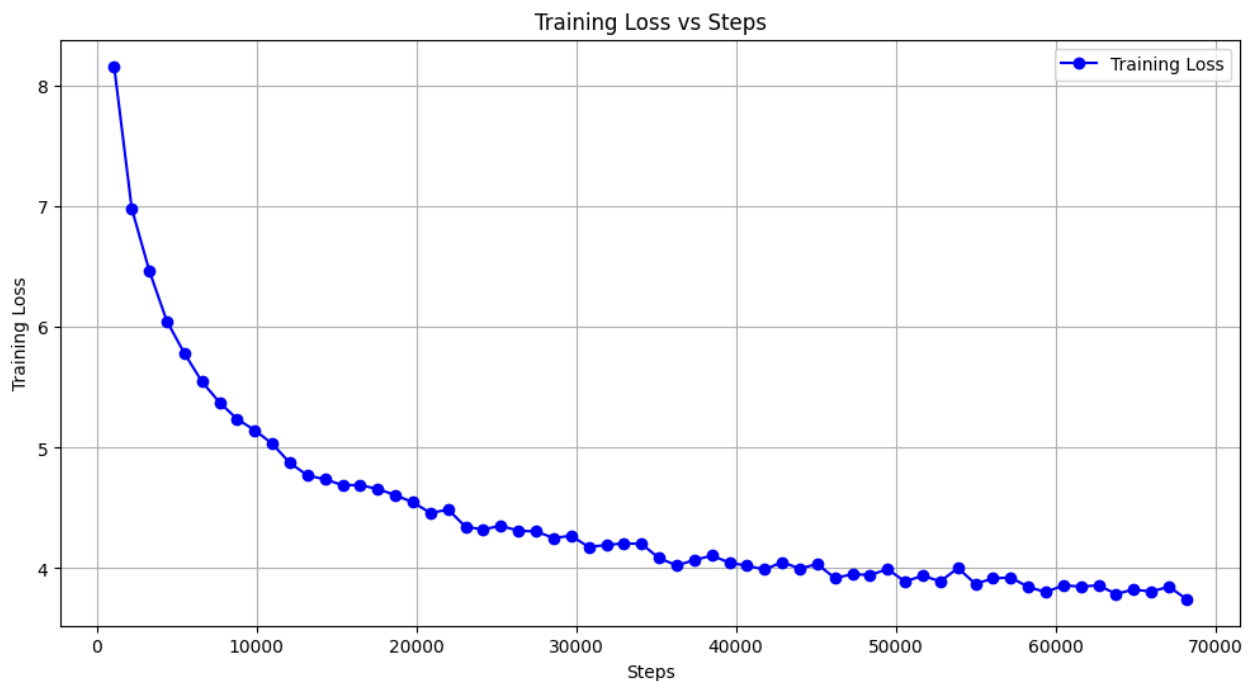
Number of parameters = 45.09 M

Tokenizer

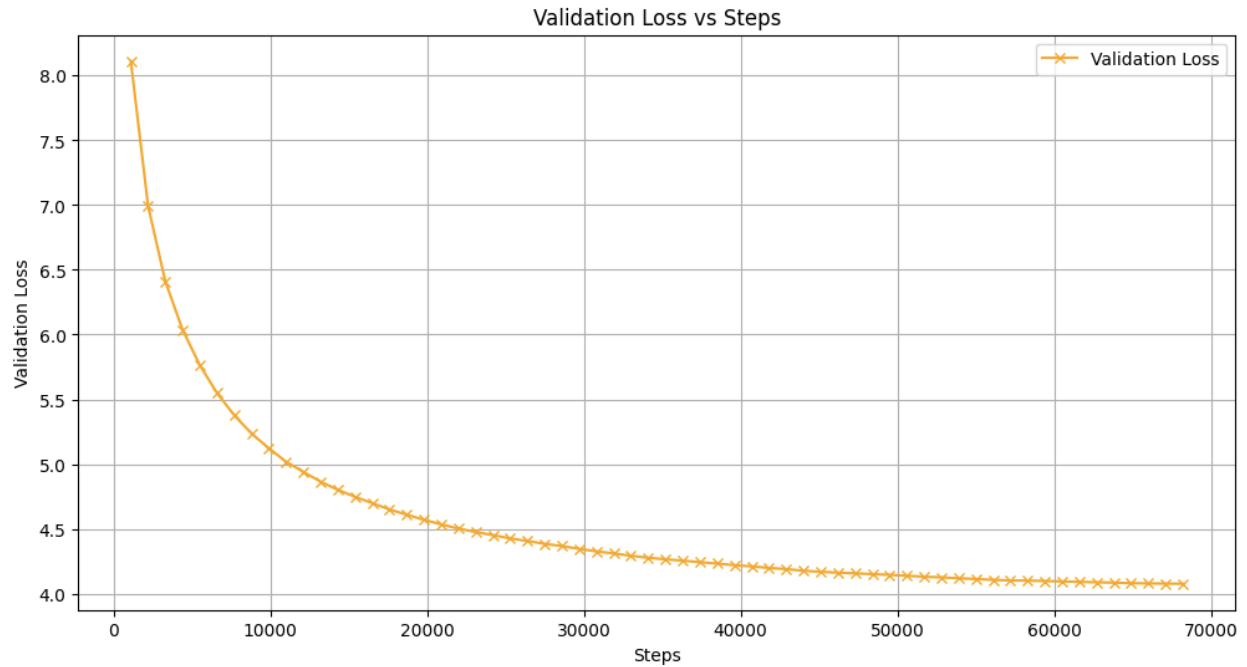
- Vocab Size: 32768
- Type: Sentence BPE

Model Evaluation:

Training Loss vs steps

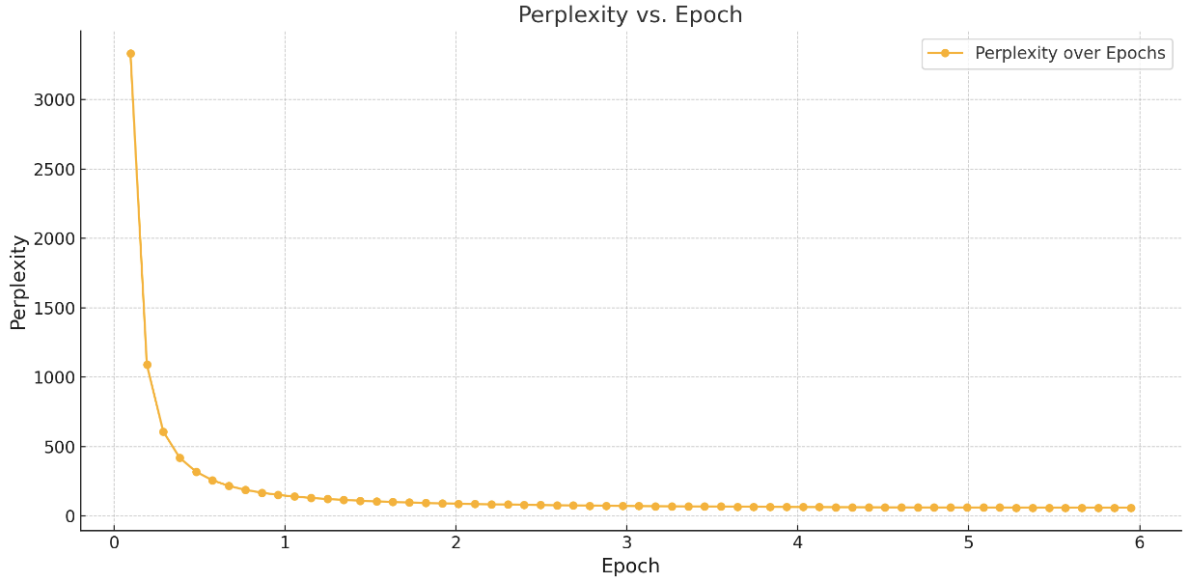


Evaluation loss vs steps



Perplexity

- **Mean Perplexity:** 172.26
- **Median Perplexity:** 71.60
- **Standard Deviation:** 432.98 (suggests high initial variation in perplexity)
- **Minimum Perplexity:** 58.92
- **Maximum Perplexity:** 3333.4



Link to Perplexity data: https://github.com/Nimitt-nim/NLP_Assignment2/blob/main/model1_perplexity_log.csv

Perplexity Matrix

Epoch	Perplexity
0	3333
1	140
2	87
3	71
4	64
5	60
6	58

Model Evaluation: Generation

Prompt 1: তোমার নাম কি?

Translation : what is your name?

Output : তোমার নাম কি?"আমরা বিশ্বাস করি না বলে আমি আপনাদের কাছে যে আমার কোনো কথাবার্তা বললে ভালো হয়, আপনারা তাদের সমর্থন ও কৃতজ্ঞতা জানেন।"খোঁজ নিয়ে জানতে চাইলে তিনি বলেন, 'আমি মনে করি যে, আমরা আপনাদের প্রতি আমাদের সম্মানিত।'ঢাকা ট্রিবিউন: স্যারও তো কথা বলতে পারব না। কারণ, তিনি একজন উপপরিদর্শক। তারা (গণমাধ্যম সমকাল) জানান, 'এই বিষয়ে আমি কিছু জানি না।"প্রসঙ্গত, ২০১৬ সালে ঢাকা বিশ্ববিদ্যালয় থেকে স্নাতক

পাস করেন তিনি। বর্তমানে তিনি ইউনিভার্সিটি অব লিবারেল আর্টস অ্যান্ড টেকনোলজি (আইসিডিডিবি) এবং তার সহযোগী বিভাগের এক্সিকিউটিভ প্রফেসর হিসেবে কাজ করেছেন। এরপর তাকে নিয়োগ দেওয়া হয়েছিল।

Transalation : What's your name? "It's better if I speak to you because we don't believe, you know their support and gratitude." 'Dhaka Tribune: Sir, I can't talk. Because, he is a sub-inspector. They (media Samakal) said, 'I don't know anything about this.' Incidentally, he graduated from Dhaka University in 2016. Currently, he has worked as an executive professor at the University of Liberal Arts and Technology (ICDDB) and its affiliated departments. He was appointed.

Prompt 2 : বর্ষার প্রথম দিনের গল্প লেখো

Translation : Write the story of the first day of monsoon

Output : বর্ষার প্রথম দিনের গল্প লেখো "এটা তার মামার সঙ্গে বলছি তুমি কী করতে?"অভিনয় দেখেছি কি না আমিও কথা বলতে পারিনি। তখন বলতে হয়, আমাদের মনের প্রতি আমার আগ্রহ নেই।আমার বয়স, আমি আজ একজন মা হতে যাচ্ছি, তিনি আরও অভিনয় করবেন যে এটা তো হলো না। আমাকে বলে মনে হয়, তোমার কোনো অনাকাঙ্ক্ষিত ঘটনা যদি আর ঘটবে না, হয়তো আমি সবসময় চাই।"তিনি আরো বলেন, "আমার ছোট ছেলে-মেয়ের সঙ্গে আমার সম্পর্কও খুব কম। বিয়েতে আমি একনজর নিয়ে নতাম, বড় মেয়েটি আমার কাছে জানতে চেয়েছি। সে কারণে আমার স্ত্রী ও মেয়েকে বিয়ে করে। আমি তাদের বাবা-ছেলে হয়ে গিয়েছিলাম। তবে আমি বিয়ে করে দিই।"এ বছর শম্ভু ও তার বাবা-মা ছাড়া বেশ কিছু পরিবারের কেউ নন। তারা যখন বাবা সন্তান নিয়ে কথা বলতে চান, তখন তার পরিবার অনেক কষ্ট পেয়ে গেলেন। কিন্তু সে কারণে বাবার জীবন শেষ পর্যায়ে কাটিয়ে দিল। এখন তাকে ভালোবেসেও বসতে বাধ্য করেছে। কারণ এটি একটি সুন্দর মুহূর্ত ছিল। তার বিশ্বাস সন্তান

Translation : Write the story of the first day of monsoon "What are you doing with this uncle?" Whether I saw the act or not I could not speak. Then I have to say, I am not interested in our mind. My age, I am going to be a mother today, he will act like it didn't happen. It seems to me, if you don't have any unwanted events, maybe I always want." I did not take a look at the wedding, the eldest daughter asked me to know. That's why he married my wife and daughter. I became their father and son. But I get married." This year there is no one from several families except Shambhu and his parents. When they wanted to talk to the father and the child, his family suffered a lot. But because of that the father's life was spent at

the last stage. Now he is forced to sit in love. Because it was a beautiful moment.
His faith is a child

Prompt 3 : একটি প্রেমের কবিতা লেখো

Translation : Write a love poem

Output : একটি প্রেমের কবিতা লেখো...অধ্যক্ষ এবং এ বিষয়ে তার কাছে কিছু বলবেন না, সে সব সময়ই কথা শুনতে হবে। আমি সেই সঙ্গেও আছি।"এর আগে, গত নভেম্বরে যুক্তরাজ্যের সেন্ট্রাল হাসপাতালের শিশু বিভাগে একটি মেডিকেল বোর্ড গঠন করা হয়। এতে একজন ডা. ইমনের বিরুদ্ধে করা অভিযোগগুলো হলো- "একজন চিকিৎসক করোনাভাইরাসে আক্রান্ত হয়েছেন; যেখানে চিকিৎসা শেষে তার জন্য প্রয়োজনীয় সহায়তার ব্যবস্থা করতে হবে। এর আগেও হাসপাতালগুলোকে কোভিড পরীক্ষা করাতে হয়েছিল। কিন্তু হাসপাতাল থেকে চিকিৎসকদের ছাড়পত্র দেওয়া হয়েছিল বলে মনে হচ্ছে, তা স্থগিত রাখা হয়েছে।"

Translation : Write a love poem...the principal and don't say anything to him about it, he must always listen. I am also with that." Earlier, a medical board was formed in the children's department of the United Kingdom's Central Hospital last November. A Dr. The allegations made against Eamon are – "A doctor is infected with coronavirus; Where necessary support should be arranged for him after the treatment. Earlier too hospitals had to conduct covid tests. But the doctors' clearance from the hospital seems to have been put on hold."

Prompt 4 : সোশ্যাল মিডিয়ার প্রভাব কি?

Translation : What are the effects of social media?

Output : সোশ্যাল মিডিয়ার প্রভাব কি?এমন প্রশ্নই করা হলো, 'সংবিধান মেনে চলব' এর নিন্দা এবং উখিয়া ও গ্রেফতারের মধ্য দিয়ে। এই ধরনের অপপ্রচার বা সহিংসতা ছড়িয়ে পড়ার কারণে আমাদের দেশগুলোকে এ রকম পরিস্থিতির মুখোমুখি হতে হচ্ছে। তাই আমি আশা করি আপনাদের ঐক্যবদ্ধ হওয়ার কোনো বিকল্প নেই- এমনটাই মনে হয়েছে যে, আপনি যদি আপনার আচরণ নিয়ে বিভ্রান্তি ছড়াতে পারে। এটা আসলে ভুলতে পারে না। তাহলে অবশ্যই এসব বিষয়গুলো আমাদের নিয়ন্ত্রণে আছে। তবে আমার মতে, আপনার যে বক্তব্য, সেটি আমি জানি না। তাহলে আমরা আইনটির ওপর প্রভাব ফেলতে পারি।

Translation : What are the effects of social media? It even has a negative effect in some cases. Because of this there is a question about the comments that will come out."

Prompt 5 : জলবায়ু পরিবর্তন মোকাবেলায় আমাদের করণীয় কি?

Translation : What can we do to combat climate change?

Output : জলবায়ু পরিবর্তন মোকাবেলায় আমাদের করণীয় কি? আমরা তা মেনে চলছি এবং বিশ্ব স্বাস্থ্য সংস্থা ও সামাজিক উন্নয়ন কর্মসূচির মতো কার্যকর ভূমিকা পালন করছে। "সংস্থাটি আরও জানায়, "বাংলাদেশ সরকার জলবায়ু পরিবর্তনের প্রভাবে এই অঞ্চলের মানুষের সংখ্যা কম বৃদ্ধি পাচ্ছে। আমাদেরকে এগিয়ে আসার লক্ষ্যে প্রয়োজনীয় পুষ্টি জোগাতে পদক্ষেপ নিতে হবে। করোনাভাইরাসের এ ধরনের বৈশ্বিক বিপর্যয়করোনাভাইরাস পরিস্থিতি যাতে নিয়ন্ত্রণে না না হয়, সেজন্য জাতীয় অভিযোজন পরিকল্পনা প্রণয়নের উদ্যোগ নেয়া উচিত।"

Translation : What can we do to combat climate change? We are complying with it and playing an effective role like the World Health Organization and social development programs." The organization further said, "Bangladesh government is reducing population growth in the region due to the impact of climate change. We need to take steps to provide the nutrients we need to move forward. Initiatives should be taken to formulate a national adaptation plan so that this kind of global disaster of coronavirus does not get out of control."

Prompt 6 : কৃষকরা অনেক সমস্যার সম্মুখীন হয়

Translation : Farmers face many problems

Output : কৃষকরা অনেক সমস্যার সম্মুখীন হয় না। "তিনি আরও বলেন, "এখন পর্যন্ত আমরা প্রায় ১০ লাখ মানুষ বিদ্যুৎ উৎপাদন করতে সক্ষম হই।"

Translation : Farmers do not face many problems." He added, "So far we have been able to generate electricity for about 1 million people."

Prompt 7 : অনেক উন্নয়ন কর্মকান্ড যেমন

Translation : Many development activities viz

Output : অনেক উন্নয়ন কর্মকান্ড যেমন হয় তেমনি তাদেরও হয়রানি করা হয় না। "তিনি বলেন, "এখন পর্যন্ত আমাদের দেশে প্রায় ১০ লাখ মানুষ বসবাস করছে। এর মধ্যে প্রায় ১০ লাখ মানুষ বসবাস করছে।"

Translation : As with many development activities, they are also not harassed." He

said, "Until now, about 1 million people are living in our country. About 1 million people live in it."

Prompt 8 : বঙ্গবন্ধু শেখ মুজিবুর রহমান

Translation : Bangabandhu Sheikh Mujibur Rahman

Output : বঙ্গবন্ধু শেখ মুজিবুর রহমান বিজ্ঞান ও প্রযুক্তি বিশ্ববিদ্যালয়ের উপাচার্য অধ্যাপক ড. মো. আখতারুজ্জামানের সভাপতিত্বে অনুষ্ঠিত সভায় আরও বক্তব্য রাখেন বিশ্ববিদ্যালয়ের উপাচার্য অধ্যাপক ড. মো. আখতারুজ্জামান।

Translation : Professor Dr. Bangabandhu Sheikh Mujibur Rahman University of Science and Technology Vice-Chancellor. Md. Akhtaruzzaman also spoke at the meeting chaired by the Vice-Chancellor of the University Dr. Md. Akhtaruzzaman.

Prompt 9 : বাংলাদেশের ক্রিকেট দলের সাম্প্রতিক পারফরম্যান্স কেমন

Translation : What is the recent performance of the Bangladesh cricket team?

Output : বাংলাদেশের ক্রিকেট দলের সাম্প্রতিক পারফরম্যান্স কেমন?"আগামী বছরের শুরুতেই বাংলাদেশ ক্রিকেট বোর্ডের (বিসিবি) সভাপতি নাজমুল হাসান পাপন ও সাধারণ সম্পাদক হিসেবে দায়িত্ব পালন করেন।

Translation : How is the recent performance of the Bangladesh cricket team?" Bangladesh Cricket Board (BCB) president Nazmul Hasan Papon served as general secretary at the beginning of next year.

Prompt 10 : একটি সূর্যাস্তের বর্ণনা দাও

Translation : Describe a sunset

Output : একটি সূর্যাস্তের বর্ণনা দাওসোনার সময় এই ঘটনা ঘটে বলে ধারণা করা হচ্ছে।

Translation : Describe a sunset that is supposed to occur during golden hour.

Model 2

Link to Model: <https://www.kaggle.com/models/nimittnimizz/bengali-llama-87m>

Configuration

```
LlamaConfig {
  "attention_bias": false,
  "attention_dropout": 0.0,
  "bos_token_id": 1,
  "eos_token_id": 2,
  "head_dim": 64,
  "hidden_act": "silu",
  "hidden_size": 512,
  "initializer_range": 0.02,
  "intermediate_size": 1024,
  "max_position_embeddings": 512,
  "mlp_bias": false,
  "model_type": "llama",
  "num_attention_heads": 8,
  "num_hidden_layers": 24,
  "num_key_value_heads": 2,
  "pretraining_tp": 1,
  "rms_norm_eps": 1e-06,
  "rope_scaling": null,
  "rope_theta": 10000.0,
  "tie_word_embeddings": false,
  "transformers_version": "4.46.2",
  "use_cache": true,
  "vocab_size": 32769
}
```

Number of Parameters = 87.05 Million

Discussion

- Model1 shows better performance in the evaluation

References

- https://github.com/AamodThakur/NLP_Pre_Training

Contribution

Nimitt

- Trained Model 2
- Trained Tokenizer 6
- Documentation

Pratham Sharda

- Trained Model 1
- Trained tokenizer 2
- Documentation

Pranjal Gaur

- Trained model 3
- Trained tokenizer 4
- Documentation

Shubham Aggarwal

- Trained model 3
- Trained tokenizer 1

Harshit

- Trained tokenizer 5
- Documentation

Chirag Patel

- Trained Tokenizer 3
- Documentation