BengaliLLama: An Instruction Following LLaMA Model for Bengali

Anonymous ACL submission

Abstract

In the field of Large Language Models (LLMs), significant advancements have predominantly focused on a limited set of languages, raising concerns in linguistically diverse regions such as India, where a wide array of regional languages are spoken, and the majority of individuals communicate in native languages other than English. Addressing this limitation, our study introduces BengaliLlama, a model tailored for Bengali, the world's seventh most widely spoken language. This research leverages a dataset of 252K Bengali instructions, translated and manually validated from various open-source resources, and employs the LoRA architecture and LLaMA for fine-tuning. The resulting BengaliLlama model demonstrates enhanced proficiency in processing and responding to instruction-based queries in Bengali. The study discussed comprehensive evaluations that will motivate various Indic Model studies in the future. BengaliLlama will be made available for research and non-commercial use, contributing to the broader goal of creating more linguistically diverse and accessible AI technologies.

1 Introduction

Since the development of pre-trained language models (Devlin et al., 2018; Liu et al., 2019), Natural Language Processing (NLP) research has achieved significant results for several different NLP tasks with specific fine-tuning (Laskar et al., 2022, 2020). There are existing Bengali pre-trained models such as BanglaBERT (Bhattacharjee et al., 2021), BanglaT5 (Bhattacharjee et al., 2023) and these models can implement tasks such as Question Answering, NLI (Natural Language Inference), NLG (Natural Language Generation). But to utilize these pre-trained models, we need large annotated Bengali datasets. However, in the NLP community, Bengali is a low-resource language and developers often face the challenges of not having large annotated datasets despite Bengali being the seventh most spoken language in the world (Sen et al., 2022).

In response to the challenges, our research mainly focuses on using Bengali as the primary language to interact with the LLM. The development of LLM in this language based on instruction sets would help develop chatbots and solve few-shot learning tasks. This paper introduces the BengaliLlama model, developed through the fine-tuning of the Bengali instruction set using Low-Rank Adaptation (LoRA) (Hu et al., 2021). Additionally, we propose a benchmarked dataset for evaluating Bengali LLMs, contributing a valuable resource to the field. Our contributions include:

- A fine-tuned open source, Bengali LLama
- One of the largest human-validated Bengali Native instruction sets of 252K instructions.
- A comprehensive evaluation study and a new benchmark of 1428 samples.

2 Dataset

The Bengali Llama utilizes a dataset of 252K Bengali instructions comprising of *a)* 152K Instructions translated from English using the AI4 Bharat Team's IndicTrans Model[1] *b)* Manually curated additional 100K samples collected from Bengali school textbooks of Class 1 till 12th for subjects like Science, Geography, History, Maths, Computer Science, Physical Education; literary work of Tagore, Ray, Bankim Chandra, etc; native folklores, food recipes, government websites, local news articles and online blogs. We developed a

[1] https://github.com/AI4Bharat/indicTrans

langchain[2] based pipeline to extract the text and structure into an instruction set. The pipeline will be released as an open-source tool by our team along with the database of the resources collected. The statistics of the dataset are shown in table 1.

Dataset	Size
Alpaca (Taori et al., 2023a)	60,402
Dolly (Conover et al., 2023)	54,456
GPTeacher[3]	9,111
GPT teacher instruct	9,987
Hard code Q&A	18,194
Manually Curated	98,146

Table 1: Details of the data used in the instruction fine-tuning stage

2.1 Human Validation

To comprehensively assess translation quality, we manually evaluated instructions translated by the IndicTrans model. Two skilled annotators (8) carried out this evaluation, each independently assigning labels to every translated segment in the sample. The purpose of using two annotators was to ensure a balanced and unbiased evaluation of the translation quality. Each segment was evaluated and labeled under five distinct categories based on translation accuracy (Parida and Bojar, 2018):

- **Flawless (F)**: Translations without any errors.
- **Good (G)**: Generally accurate translations need minor corrections.
- **Partly Correct (PC)**: Translations are accurate in parts but with some mistranslations.
- **Ambiguity (A)**: Cases where the meaning of a word was misunderstood.
- **Incomplete (I)**: Correct translations but truncated or missing some content words.

We employed a mathematical approach to quantify the evaluation results to average the scores across both annotators for each category. The average score for a category C was calculated as follows:

$$\text{Avg Score}_C = \frac{\text{Score}_{1,C} + \text{Score}_{2,C}}{2} \quad (1)$$

The Human Validation Summary can be seen in Table 2.

This manual annotation process and a mathematical averaging approach ensured a thorough and unbiased translation quality assessment. It can be used to validate the effectiveness of the IndicTrans model and as a support to the language model training for Bengali.

Cat.	Score1	Score2	Avg Score	(%)
F	56588	54472	55530	36.5
G	53728	56432	55080	36.5
PC	17948	16232	17090	10.3
A	13388	14712	14050	8.5
I	10348	10152	10250	8.2

Table 2: the average score represents the mean value of scores assigned by two annotators for each category. The percentage indicates the proportion of the total dataset that falls into each category.

3 Model Building

We adopted Low-Rank Adaptation (LoRA) for model building, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture (Hu et al., 2021). We used Large Language Model Meta AI (LLaMA) (Touvron et al., 2023) as the foundation model for fine-tuning. Due to the smaller size, LLaMA requires fewer computing resources, and we used LLaMA 7B for fine-tuning, which is trained on one trillion tokens with a majority of data in English.

We followed a methodology that was employed in Stanford Alpaca (Taori et al., 2023b) to implement self-guided fine-tuning for training the instruction-following model. Each instance comprises an instruction and a corresponding output.

4 Experimental Setting and Training

We trained the model on Nvidia A100 PCIE GPU with 40 GB. The model was trained for a total of five epochs, which took approximately four days to complete. The hyperparameters eventually used for the fine-tuned model are shown in Table 3.

5 Inference

5.1 Automatic Evaluation

To comprehensively evaluate BengaliLlama, we followed the methodology outlined in (Kabir et al., 2023). This involved studying various NLP tasks, as summarized in Table 4 of the mentioned work. We collected the necessary datasets and applied the prompt technique detailed in the study to compare

[2] https://www.langchain.com/
[3] https://github.com/teknium1/GPTeacher

Hyper Parameter	Value
Batch Size	128
Learning Rate	$3e^{-4}$
Epochs	5
Cutoff Length	256
Weight_Decay	0.001
Warmup_Rate	0.1
LR_Scheduler	linear
Lora r	16
Lora Target Modules	(q_proj, k_proj, v_proj, o_proj)

Table 3: Training Hyperparameters

our models with ChatGPT, Base Llama2-7B, Claude-2, and Mistral 7-B. The results of these comparisons are reported in Table 4.

In summarization and paraphrasing, ChatGPT scored highly due to its generation of responses with a higher word average than other models, which led to its highly descriptive outputs. In this case, Base Llama2-7B's performance was significantly lower as it returned results in English translation of Bengali, which rendered the evaluation insignificant. Our finetuned model demonstrated excellent capabilities in generation tasks, achieving higher performance than Mistral and BaseLLama2.

In Question-Answering, our model did not perform well due to its inability to match the golden labels exactly. On closer observation, we found that our model provided outputs that primarily aligned well with the context; however, their generated answers were wrong or included incorrect or redundant details. Similar observations were made in some cases of correct answers as well when unnecessary additional details accompanied them.

In NLI, sentiment analysis, and text classification, our model outperformed Base Llama2 7B and Mistral 7B. We observed that larger models such as ChatGPT and Claude-2 often exhibit bias towards expressing a particular opinion polarity (contradiction, entailment) when dealing with logical relationships in Bangla. Our model, on the other hand, demonstrated proficiency in capturing polarities in the classification task. We also noted that the induction of local context throughout the instruction set enhanced our model's capability to avoid the general bias that ChatGPT suffers from due to its extensive knowledge base. The specific knowledge base seems to be impactful.

Our team also designed a sample dataset of 428 samples extracted from the various literary resources of Bengali literature featuring diverse styles: **Rabindra Rachanabali** embodies lyrical depth and symbolism, **Nazrul Geeti** reflects revolution and patriotism, **Saratchandra's realism** focuses on societal struggles, **Bankim Chandra's romanticism** celebrates nature and patriotism, **Humayun Ahmed's modernism** explores human emotions, and **Satyajit Ray's simplicity** conveys profound narratives. Folk literature preserves Bengali customs and values through traditional stories, songs, and poems. It provides us with a great way of evaluating language models' capacity to be familiar with the core native ideologies and styles. This dataset will also be pivotal in critically evaluating LLMs prepared in the future. We have analyzed the results using the aforementioned metrics in Table 5.

5.2 Human Evaluation of Model Reasoning and Understanding

The Human Evaluation task is conducted here to assess the comparison between BengaliLLama, ChatGPT, and BaseLLama2 Models. Two native Bengali speakers have participated in this task; the inter-annotator agreement is mentioned in Appendix. To support the automatic evaluation, we have conducted a human evaluation of the model by these native Bengali speakers. We have measured the performance based on three metrics namely correctness, perplexity, and readability. A tabular representation is provided in the appendix with details about the scoring. From 1 to 5, scoring is used where 1 being the lowest and 5 being the highest. We have observed that BaseLLama2 provided mostly English outputs and in some cases, the output came out to be out of context; hence, we have chosen to score BengaliLLama and ChatGPT outputs as per three metrics. We have tested these metrics of outputs using TruthfulQA (Lin et al., 2022) (dataset consisting basic general knowledge questions and answers), LogiQA (Liu et al., 2020) (dataset consisting basic mathematical questions and answers), MMLU (Hendrycks et al., 2021) Logical Fallacies (dataset of logical fallacies questions and answers), MMLU Philosophy (dataset of philo-

	XL-Sum (AS)	SQuAD_Bangla (QA)	Indic Para (PP)	BNLI (NLI)	SNAC (TC)	IndicSent (SA)	SentNoB(SA)
Model	*R1/R2/RL*	*EM/F1*	*Bleu*	*Acc*	*Acc*	*Acc*	*P/R/F1*
ChatGPT	27.11/8.07/20.86	44.85/78.67	2.81	52.71	18.36	90.20	57.70/54.56/53.17
Llama-2-7b	4.51/0.17/1.42	31.73/67.95	0.01	42.37	14.47	69.16	48.39/48.49/48.43
Claude-2	21.97/6.06/17.55	49.92/79.04	1.89	32.20	20.76	88.48	53.28/54.38/52.79
Mistral7B	20.53/5.75/15.63	42.11/76.20	1.92	36.86	17.49	86.34	53.23/54.01/53.97
BengaliLLama	**22.18/6.19/18.03**	**39.23/77.67**	**2.23**	**48.88**	**18.77**	**83.63**	**55.22/54.96/54.74**

Table 4: Comparative analysis of LLMs across various NLP tasks in Bengali; Abstractive Summarization (AS), Question Answering (QA), Paraphrasing (PP), Natural Language Inference (NLI), Text Classification (TC), and Sentiment Analysis (SA), where evaluation metrics are Exact Match (EM), Accuracy (Acc.), Precision (P), Recall (R), and F1 Score (F1).

Tasks	**Bert Score**		**Rouge**					
	Precision	**Recall**	**F1**	**rouge1**	**rouge2**	**rougeL**	**rougeLSum**	**Bleu**
BengaliLlama	0.38	0.35	0.37	0.4	0.29	0.3	0.33	28.3
ChatGPT	0.41	0.39	0.40	0.44	0.29	0.36	0.37	33.4
Mistral 7B	0.33	0.32	0.33	0.3	0.20	0.25	0.25	20.1
Claude-2	0.35	0.33	0.35	0.33	0.26	0.29	0.3	22.5

Table 5: NLG Metrics Comparison on the manual created literature dataset

Dataset	**Size** (number of questions)
TruthfulQA	200
LogiQa	200
MMLU Logical Fallacies	100
MMLU Philosophy	250
MMLU Jurisprudence	250

Table 6: Manual Evaluation Dataset Statistics

sophical questions and answers) and MMLU Jurisprudence (dataset of legal philosophy). We have taken 1000 samples of these datasets translated and validated by annotators.

- Correctness - For the TruthfulQA dataset, it is observed that ChatGPT provided out-of-context answers in 40% cases when the context is not provided. However, when context is provided to ChatGPT, it provided correct answers. The correctness of ChatGPT is scored as 3.25 for this dataset; Whereas, BengaliLLama provided correct answers always within the context whether the context is provided or not. For other datasets, BengaliLLama provided wrong outputs, hence, the average scoring is provided 2.6 for BengaliLLama and 3 for ChatGPT.
- Perplexity - For TruthfulQA dataset, both BengaliLLama and ChatGPT performed well (scored 1 for both models). For LogiQA dataset, ChatGPT provided clear mathematical explanations and outputs whereas, BengaliLLama provided confusing Bengali words and wrong outputs. For Logical Fallacies, Philosophy, and Jurisprudence, BengaliLLama provided outputs with wrong spellings, bad sentence construction and meaningless answers. however, ChatGPT provided answers without spelling mistakes or sentence construction mistakes but all of them are out-of-context answers.
- Readability - We have obtained readable outputs from BengaliLLama and ChatGPT models for all the evaluation datasets. However, for MMLU Jurisprudence dataset, BengaliLLama provided non-readable outputs. Therefore, an overall, average scoring was provided 4.7 for BengaliLLama and 5 for ChatGPT.

6 Conclusion

This study signifies a significant leap forward in NLP for Bengali, a language that has historically been underrepresented in the field of large language models. The creation and refinement of the BengaliLLama model, leveraging the LLaMA architecture with LoRA optimizations, represent a crucial development in addressing Bengali's linguistic and cultural nuances. Releasing a 252K validated Bengali instruction set significantly contributes to research in low-resource languages. This dataset enhances the depth of research in Bengali NLP and serves as a valuable resource for the broader linguistic community.

7 Limitations and Future Work

This study, employing metrics like ROUGE, BLEU, and BERTScore, acknowledges key limitations in its evaluation methodology. An over-reliance on automated metrics is evident; while they offer quantitative rigor, they may not capture the model's nuanced handling of the Bengali language, particularly in cultural and idiomatic expressions. The lack of human evaluators, especially those fluent in Bengali, is a significant gap, as automated metrics often overlook the subtleties of natural languages.

Future research should embrace a comprehensive, multi-faceted evaluation approach. Integrating qualitative assessments by native Bengali speakers is crucial for a deeper understanding of the model's handling of cultural nuances and idiomatic expressions. Expanding the range of linguistic tasks—including translation, sentiment analysis, conversational AI, and other NLG tasks—will provide a richer understanding of the model's versatility. For authentic and practical model evaluation, employing diverse datasets that capture the full breadth of Bengali's linguistic diversity, including its dialectal and colloquial varieties is essential. Comparative studies with other models, coupled with real-world application testing and detailed error analysis, will offer invaluable insights into the model's practical utility and areas for improvement.

Ethics Statement

We do not envisage any ethical concerns. The dataset does not contain any personal, or personally identifiable, information, the source data is already open source, and there are no risks or harm associated with its usage.

References

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, and Rifat Shahriyar. 2023. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 714–723.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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*.

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2023. Benllmeval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.

Md Tahmid Rahman Laskar, Xiangji Huang, and Enamul Hoque. 2020. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5505–5514.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shantipriya Parida and Ondřej Bojar. 2018. Translating short segments with NMT: A case study in English-to-Hindi. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 249–258, Alicante, Spain.

Ovishake Sen, Mohtasim Fuad, Md Nazrul Islam, Jakaria Rabbi, Mehedi Masud, Md Kamrul Hasan, Md Abdul Awal, Awal Ahmed Fime, Md Tahmid Hasan Fuad, Delowar Sikder, et al. 2022. Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning based methods. *IEEE Access*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023a. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023b. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

8 Appendix

8.1 Inter-annotator Agreement

We have recruited Bengali speakers as the team of annotators. The team was formed including four female and three male annotators. Among the team, two of them validated the entire dataset and each of the members of the annotation and validation team have at least an undergraduate degree. They all reside in different parts of West Bengal. We have provided annotation guidelines as follows:

- Remember to read the Bengali typing rules. Before starting test it once and report for any issues.
- The annotator must be a native speaker for the annotation task.
- Look at the image before annotating.
- Try to understand the task i.e. translate the questions and answers in the respective language.
- Do not use any Machine translation system for annotation.
- Do not enter dummy entries for testing the interface.
- Data will be saved at the backend.
- Press the Shift Key on the virtual keyboard for complex consonants.
- Contact the coordinator for any clarification/support

8.2 Translation of instructions into Bengali

8.2.1 Indic Trans

This Transformer-4x NMT system, boasting approximately 434M parameters, was trained on the extensive Samanantar dataset, which is pivotal for Indic languages. IndicTrans enhances translation efficiency by standardizing all Indic data into the Devanagari script, fostering improved lexical consistency and subword vocabulary compactness, which benefits languages like Bengali. Its effectiveness is demonstrated by superior performance in benchmarks such as WAT2021, WMT, UFAL, and PMI, surpassing open-source counterparts and competing with major commercial systems.

8.2.2 Translation Validation Score Calculation

For instance, if Annotator 1 labeled 80 segments as 'Flawless' and Annotator 2 labeled 75 segments likewise, the average score for the 'Flawless' category would be $(80 + 75)/2 = 77.5$. This averaging method was applied to all categories, providing a balanced and statistically sound evaluation of the dataset's translation quality.

8.3 Training Setup Summary

A batch size of 128 and a learning rate of 3e−4 were employed. To prevent overfitting, a weight decay of 0.001 was applied. The training process also incorporated a warmup rate of 0.1 to increase the learning rate gradually. The learning rate scheduler followed a linear function. The model architecture utilized a Lora r of 16 and targeted specific modules, including q_proj, k_proj, v_proj, and o_proj. Additionally, a cutoff length of 256 was used to limit the input length during training. These experimental settings were carefully selected to ensure an optimal balance between computational resources and model performance. The training and evaluation loss are shown in Fig. 1 and Fig. 2.

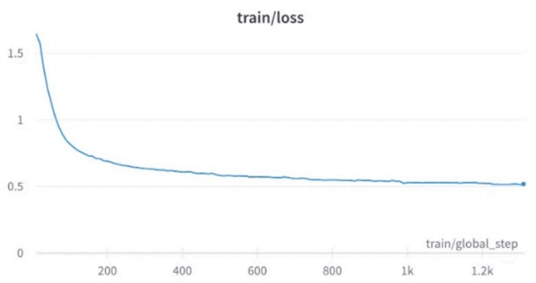

Figure 1: Training loss

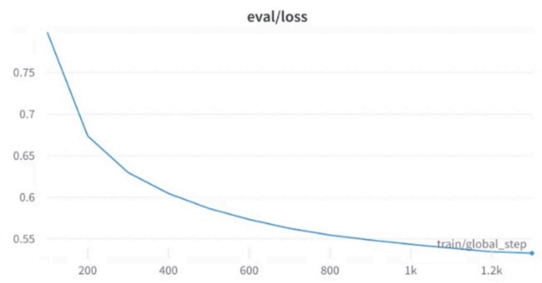

Figure 2: Evaluation loss

8.4 Inference

8.4.1 Text Generation Setup

The decoding process of LLMs plays a critical role in determining the quality and diversity of the generated text. In our experiments, we use the following decoding hyperparameters:

- *Size of the context*: We establish the context size as 2048, determining the maximum number of tokens that the model can take into account simultaneously during the text generation process.
- *Maximum sequence length*: We impose a constraint on the generated sequence length, limiting it to 512 tokens to ensure that the outputs remain focused and closely related to the input prompt.
- *Temperature*: We set the temperature to 0.2, regulating the level of randomness in the sampling process. Lower values make the model produce more focused and deterministic outputs, while higher values introduce greater diversity at the expense of coherence.
- *Top-k sampling*: For each step, we adopt Top-k sampling with a value of k = 40, whereby the model selects the subsequent token from the top 40 most probable options. This introduces an element of randomness and diversity in the generated text.

8.4.2 Inference Screenshots

Some of the interface screenshots are shown here with their output obtained from the BengaliLLama model. The images in figure 3, 4, 5, 6 are showing the questions and answers.

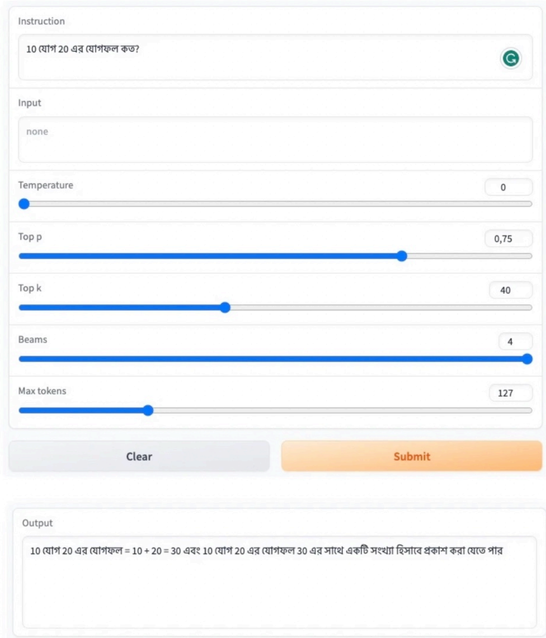

Figure 3: Sample Inference 1. Question: "What is the sum of 10 plus 20?" Answer: "The sum of 10 plus 20 = 10 + 20 = 30 and the sum of 10 plus 20 can be expressed as a number with 30"

Figure 4: Sample Inference 2. Question: "What are the benefits of eating an apple a day?" Answer: "Benefits of eating an apple a day Apples are a healthy and wholesome food to eat."

Dataset	Correctness*		Perplexity		Readability	
	Bengali LLama	*Chat GPT*	*Bengali LLama*	*Chat GPT*	*Bengali LLama*	*Chat GPT*
Truthful QA	5	3.25	1	1.75	5	5
LogiQA	1	5	3	1	5	5
MMLU Logical Fallacies	5	3	1	1	5	5
MMLU Philosophy	1	2.5	5	1	5	5
MMLU Jurisprudence	1	1	3.5	2	3.5	5
Average	2.6	2.95	2.7	1.35	4.7	5

Table 7: Brief human evaluation of the dataset [*Indicates (1-5) rating for each metrics - 1 being the lowest and 5 being the highest.]

Question	Answer given by our BengaliLLama Model	Answer given by ChatGPT Model	Answer given by BaseLLama2 Model
গিরগিটি কেন রং পরিবর্তন করে? (Why does a chameleon change its color?)	গিরগিটি তাদের তাপমাত্রা নিয়ন্ত্রণ করতে রং পরিবর্তন করে? (A chameleon changes its color to regulate their body temperature.)	গিরগিটি রং পরিবর্তন করে কারণ তার শরীরে থাকা কিছু একটি অসুস্থতা অথবা প্রাকৃতিক কারণ হতে পারে। (A Chameleon can change its color because of its physical condition or due to natural reasons.)	Camouflage.
১৪২.৭৬ - ১৬.৫ এর পার্থক্য খুঁজুন। (Find the difference between 142.76 and 16.5)	১৪২.৭৬ - ১৬.৫ = ১২৬.২১ এর পার্থক্য খুঁজে পাওয়া যায়। তাই উত্তরটি ১২৬.২১ হবে। (The difference between 142.76 - 16.5 = 126.21 can be found. So, the answer will be 126.21)	১৪২.৭৬ - ১৬.৫ এর পার্থক্য হলো ১২৬.২৬। (The difference between 142.76 - 16.5 is 126.26)	142.76−16.5=126.26

Table 8: Comparative analysis between BengaliLLama, ChatGPT and BaseLLama2

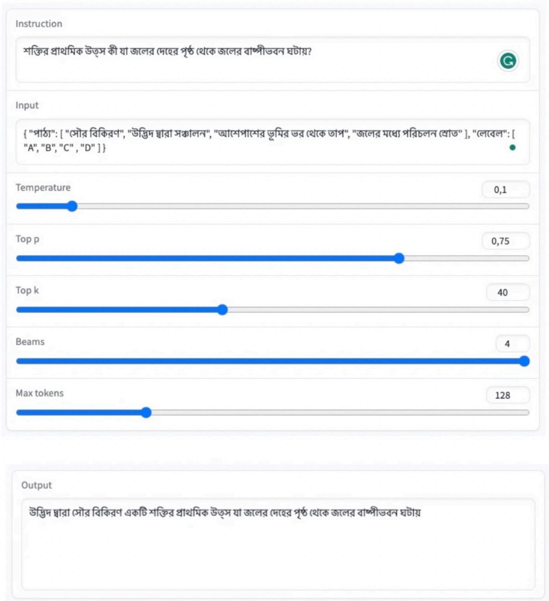

Figure 5: Sample Inference 3. Question: What is the primary energy source that causes water's evaporation from the surface of a body of water? input is, "text": ["solar radiation", "conduction by plants," "heat from surrounding land mass", "convection currents in water"], "label": ["A", "B", "C", "D"] Answer: "Solar radiation by plants is a primary source of energy that causes evaporation of water from the surface of water bodies."

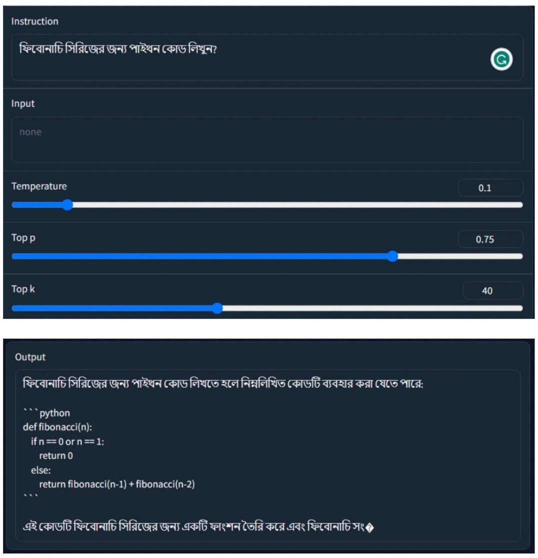

Figure 6: Sample Inference 4. Question: "Write Python code for the Fibonacci Series". Answer: "The following code can be used to write Python code for the Fibonacci Series [python code]"