PARAMANU: A FAMILY OF NOVEL EFFICIENT GENERATIVE FOUNDATION LANGUAGE MODELS FOR INDIAN LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

We present PARAMANU (which means "atom" in multiple Indian languages), a *family* of novel language models for *Indian languages*. It is a collection of auto-regressive monolingual, bilingual, and multilingual Indian language models pretrained from scratch, currently covering 10 Indian languages (Assamese, Bangla, Hindi, Konkani, Maithili, Marathi, Odia, Sanskrit, Tamil, Telugu) across 5 scripts (Bangla, Devanagari, Odia, Tamil, Telugu). The models are pretrained with a context size of 1024 on a single GPU, and are of varying sizes ranging from 13.29 M to 367.5 M parameters. We proposed a RoPE embedding scaling method that enables us to pretrain language models from scratch at larger sequence length context size on single GPU without increased GPU memory. ~~than the equivalent GPU memory.~~ We have also developed an efficient and advanced novel tokenizer with least fertility score among existing LLMs for Indian languages using a combination of BPE and Unigram that can also tokenize unseen languages written in the same script or the Roman script. We also proposed language specific tokenization for multilingual models and domain specific tokenization for monolingual language models. In order to avoid the "curse of multi-linguality" in our multilingual MPARAMANU model, we pretrained on comparable corpora by typological grouping using the same script. We proposed and performed pretraining for more than 1 epoch of training for most of our language models. From our results, we observed the language transfer phenomenon from low resource to high resource within languages of the same script and typology. We performed human evaluation of our pretrained models for open end text generation on grammar, coherence, creativity, and factuality metrics for several languages. Our Paramanu models outperformed standard and multilingual large language models (LLMs) by a large margin in performance despite being smaller in size by 64 to 20 times. We studied the impact of language specific tokenization versus language agnostic tokenization for bilingual language modeling. We also studied the impact of BPE versus Unigram tokenization for Devanagari script languages. We further created instruction-tuning datasets and instruction-tuned our pretrained models on 23,000 instructions in respective languages except Hindi, for which we used 75,000 instructions. Comparison with multilingual LLMs on various commonsense reasoning benchmarks for natural language understanding, natural language inference, and machine reading comprehension shows the advantage of our models. The performance of our Paramanu models leads to the conclusion that high quality generative language models are possible without high amount of compute power (FLOPS) and enormous number of parameters.

1 INTRODUCTION

World's new age growth arenas are non-English speakers from countries such as India, China, Indonesia, Mexico, South Africa, etc., where more people will connect to the Internet for information need without language acting as boundaries. Multi-lingualism represents freedom of expression and diversity in a country like India. According to a Guardian article[1] by Andras Kornai, "95% of

[1] http://labs.theguardian.com/digital-language-divide/

all languages in use today will never gain traction online"; this highlighting the digital language divide. Despite around 7,000 languages, current NLP technologies cover only 12%, leaving many non-English and non-European languages underserved. This, in our personal opinion, is an extremely incapacitating bias and language divide for the global digital world where the non-English, non-European language world has been left out. According to the Indian Census 2011, there are 22 official languages and more than hundred others with a sizeable number of speakers in India. Both Hindi and Bangla (Bengali), despite being the world's 5th and 6th most spoken languages respectively according to Babel[2] are still underrepresented in today's NLP technology. Multilingual NLP faces the challenges of having lack of quality benchmark datasets covering diverse languages from different language families and especially under-represented languages, and is typically referred as "low-resource" in the NLP community. Large language models (LLMs) such as GPTNeoX (Black et al., 2022), OPT (Zhang et al., 2022), LLaMa (Touvron et al., 2023), PaLM (Chowdhery et al., 2023), GPT-2-XL (Radford et al., 2019), GPT-J (Wang & Komatsuzaki, 2021), etc. have primarily focused on English and mostly European languages whereas other languages have not been given priority. Bloom (Workshop et al., 2023) is considered to be the biggest multilingual auto-regressive model that has been built till now; it has been pretrained on 45 languages including Indian languages. However, Indian languages are morphologically richer and typologically distinct than languages written in the Latin script and, hence, grouping them together without considering linguistics nuances often leads to poor performance of LLMs for low resource languages.

This work is an attempt to make language technology more accessible for Indian languages. In this work, we focus on 10 Indian languages (Assamese, Bangla/Bengali, Hindi, Konkani, Maithili, Marathi, Odia, Sanskrit, Tamil, Telugu) written in 5 distinct scripts (Assamese-Bengali, Devanagari, Odia, Tamil, Telugu) comprising of more than billion speakers in the global world. Our goal is to show that generative language models for low-resource Indian languages can be trained from scratch with limited compute and token budget, without using English corpora. We excluded English due to its linguistic differences with Indian languages in terms of typology, script, morphology and grammar, aiming to maintain language purity and typology grouping. Additionally, existing multilingual models often have an English-centric bias due to the large imbalance in data. To the best of our knowledge, the existing LLMs and multilingual LLMs struggle to generate grammatically correct and coherent sentences in Indian languages (and reasoning ability is too far) as shown in Appendix Tables 20, 26, 27, 28, 29, 30, 30, 32, 33, 35, 36, 37, 38, 39, 40. Many LLMs even generate text in even Arabic (Table 29) or Japanese (Table 39) scripts when prompted with prompts in Bengali, Hindi, and Sanskrit. We also found that LLMs including ChatGPT (Oct'23) were not able to distinguish languages of the same script such as Bengali versus Assamese. Human evaluation by annotators confirm this (Section A.5 in Appendix). Our models, based on Transformer decoders (Vaswani et al., 2017), are enhanced with improvements in the architecture. These enhancements, improvements and novelties make our models efficient, small but strong. They have been pretrained from scratch to support context size of 1024 without requiring higher physical memory on a single NVIDIA A100-PCIE-40GB GPU. Our model architecture has the ability to capture a much higher context size without requiring equivalent physical memory~~; in contrast, the LLaMa-1 model was pretrained from scratch using 2048 GPUs for a context size of 2048~~. Our models are of varying sizes and of three types: (1) monolingual language models, (2) bilingual language models, and (3) multilingual language models. Both bilingual and multilingual language models are pretrained from scratch on comparable corpora and with typological grouping of languages to avoid curse of multilinguality.

We summarize our contributions as follows.

1. We propose a RoPE (Su et al., 2022) embedding scaling method that enables us to pretrain language models from scratch at larger sequence length context size than the equivalent GPU memory. We scaled the RoPE embedding through a shrinking factor by dividing the target context length y by max_permissible_context_size_length on single GPU keeping all other hyperparameters fixed such as batch size, vocabulary size.
2. We propose a novel tokenization method using a combination of both Byte-Pair Encoding (BPE) (Sennrich et al., 2016) and Unigram (Kudo, 2018) algorithms.
3. We perform language specific tokenization for multilingual model and domain specific tokenization for monolingual language modeling.

[2]https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world

4. We perform pretraining on comparable corpora for multilingual/bilingual generative language model to handle data imbalance and curse-of-multi-linguality in multilingual language model.
5. We perform more than 1 epoch of pretraining for most our language models. From our results, we observed the language transfer from low resource to high resource within language of the same script and typology. We found that small models trained for more than 1 epoch on high quality data is better than bigger model trained for 1 epoch on not so good quality bigger corpora. We studied the impact of language specific tokenization versus language agnostic tokenization for bilingual language modeling. We also studied the impact of BPE vs Unigram tokenization for Devanagari script languages as a proof-of-concept.
6. We curated books for pretraining dataset and created an instruction-tuning dataset of 23000 instructions and instruction-tune our pretrained Bangla, ~~Hindi,~~ Marathi, Tamil, and Telugu models using 23000 instructions each in respective languages but for Hindi, we used additional 52,000 Alpaca translated Hindi (Taori et al., 2023), total of 75,000 instructions.

2 Related Work

The rise of large language models (LLMs) has led to numerous new models competing across various benchmarks, but most remain English-centric. Bloom, the largest multilingual LLM with 170 billion parameters, utilizes 40% English data and 60% from 44 other languages. Despite extensive multilingual training, many LLMs exhibit a significant English bias. Research has shown that a substantial portion of neurons in these models remains inactive; for instance, in the 66 billion parameter OPT model, over 70% of the feed-forward network (FFN) neurons in certain layers are "dead" meaning they do not activate even on diverse datasets. This sparsity in neuron activation limits the model's ability to learn and generalize across languages, particularly those that differ significantly from English (Voita et al., 2024). Consequently, even when trained on multilingual data, these models struggle to effectively process and generate text in languages beyond English, revealing a critical limitation in their design. Recent efforts to adapt English-centric models like Llama for Indian languages (Airavata (Gala et al., 2024), OpenHathi for Hindi, and Tamil) have involved extending vocabulary and fine-tuning techniques such as LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023). However, these models still exhibit a strong English bias and struggle to generate high-quality text in Indian languages. Massively multilingual models (MMTs) (Devlin et al., 2019), (Conneau et al., 2020), (Xue et al., 2021) are pretrained on large corpora but often lack alignment between languages, leading to poor transfer performance for distant languages (Lauscher et al., 2020). The "curse of multilinguality" (CoM) indicates that adding more languages can degrade per-language performance, necessitating larger model capacities and corpora. While language-specific adapters like MAD-X (Pfeiffer et al., 2020) improve performance, they do not generalize well to unseen languages. Recent advancements, such as MAD-G (Ansell et al., 2021) and BAD-X (Parović et al., 2022), focus on bilingual adapters to enhance language transfer. Our work aims to create dedicated models for low-resource Indian languages by developing language-specific generative models from scratch, emphasizing linguistic features, typology, and tailored tokenization.

3 Methodology

3.1 Dataset for Pretraining

Pretrained data was split into 95%-5% training and validation sets so that we do not lose much data for pretraining as the purpose of this work is a step towards developing pretrained generative effective language models from scratch using our novel architecture for Indian languages. Pretraining data covers web scrapped news, blogs from IndicCorp v2 (Doddapaneni et al., 2023), Bangla literature from Vacaspati (Bhattacharyya et al., 2023), Wikipedia articles, curated books of various genres, subjects, education books, magazines in respective Indian languages representing each distinct language community to cover Indian culture, rich history, and knowledge. Our pretraining corpora have no source code, scientific journals/articles, medical and engineering education books, research papers as these are generally written in English in India. Dataset details can be found in Table 19 in Appendix and data distribution can be found in the Figure 3 in Appendix. We apply the following data cleaning and preprocessing techniques as mentioned in Appendix A.4.

Models	Size	#Tokens pretrained	# Training A100 hours	Script & Family	# Speakers
Bloom	7.1B	340 B	1.08M	Multilingual	1B+
OpenHathi (Llama)	7B	1 T	N/A	Hindi (Indo-European)	692M
Sarvam	2B	4 T	N/A	Multilingual (Devanagri, Dravidian)	1B+
Paramanu-Bangla (ours)	108M	26.21 B	19.45	Bengali (Indo-European)	300M
Paramanu-Hindi (ours)	367M	66 B	239	Devanagari (Indo-European)	692M
Paramanu-Marathi (ours)	208M	28.83 B	88	Devanagari (Indo-European)	99M
Paramanu-Odia (ours)	87M	52.42 B	84.5	Odia (Indo-European)	43M
Paramanu-Sanskrit (ours)	139M	45 B	110	Devanagari (Indo-European)	0.025M
Paramanu-Tamil (ours)	208M	26.2 B	208	Tamil (Indo-Dravidian)	77M
Paramanu-Telugu (ours)	208M	39.32 B	112.5	Telugu (Indo-Dravidian)	95M
mParamanu (ours)	162M	26.2 B	118	Multilingual (Devanagari)	1B+

Table 1: Pretrained LLM tokens, training hours, script, language family and #speakers.

3.2 Dataset for Instruction Tuning

We created a dataset of 5,000 human-annotated instructions covering tasks like poem and novel writing, article summarization, grammar correction, and Q&A on topics such as climate change and healthcare in Bangla, reflecting Bengali culture and linguistics. We then used the Google Translate API (goo, 2023) to translate these instructions into Hindi, Marathi, Tamil, and Telugu. Additionally, we translated 15,000 instructions from Dolly (Conover et al., 2023) to these languages and generated 3,000 instructions using the self-instruct technique (Wang et al., 2023) in five Indian languages. Finally, we fine-tuned our pretrained models except Hindi on a total of 23,000 instructions, including human-annotated, machine-generated, and translated instructions. Since creating multilingual datasets using automatic machine translation from Bengali to other Indian languages introduce translation errors but we also performed human checks and corrected them. We found around 8% word errors on average for Hindi and Marathi using Google Translate.

3.3 Tokenization

Figure 1 shows the flowchart of our novel tokenization technique both for monolingual and multilingual settings. We performed domain adaptive tokenization for monolingual models using a combination of Byte-Pair encoding (BPE) (Sennrich et al., 2016) and Unigram (Kudo, 2018). We trained separate Byte-Pair encoding (BPE) (Sennrich et al., 2016) and Unigram tokenizers using Sentencepiece (Kudo & Richardson, 2018) module on the high quality part of the pretraining data from scratch. Then, we merge both the independent tokenizers by intersection respectively of size V‘. This is similar to merging two different data structures of same size (considering each tokenizer as a list of tokens) by intersection to remove overlapping elements. We used the merged tokenizer to tokenize the pretraning data. During pre-tokenization, NFC normalization was performed on the processed data; digits are split into individual tokens and unknown UTF-8 characters were reduced to byte granularity. For monolingual tokenization, we trained individual BPE and Unigram tokenizers on domain specific data for respective language with same vocabulary size and merged the tokenizers via merge by intersection to remove overlapping tokens and make our specialised tokenizer compact, optimized, effective and highly effective for monolingual data and also performed the same approach to tokenize the multilingual pretraining corpora by performing language specific tokenization in multi-task way where every task is a language and merged the distinct language specific tokenizers by union via intersection, i.e, removing the overlapping tokens. In this way, our mBharat tokenizer was able to learn language specific tokens based on typology and can generalize to unseen languages of the same script. mBharat tokenizer was also exposed to little amount of English high quality corpus to learn its ability to tokenize languages of the Roman script such as English. From 2, we observe that mBharat tokenizer has the least fertility score of 1.66 for languages in Assamese-Bengali, 1.25 for Devanagari (Hindi) and 1.75 for Odia script among tokenizers of LLMs like Sarvam 2B (Sarvam2B, 2024), LLama-3.1 (Dubey et al., 2024), Gemma-2 (Team et al., 2024), and GPT-4o.

3.4 Model Architecture

Multilingual mParamanu, monolingual Paramanu models (Assamese, Bangla, Hindi, Konkani, Marathi, Odia, Tamil, Telugu, and Paramanu-Sanskrit) and bilingual Konkani-Maithili models are based on transformer (Vaswani et al., 2017) based causal decoder architecture (Radford et al.,

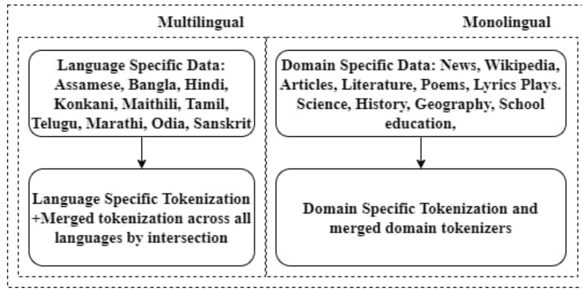


Figure 1: Tokenization technique for monolingual and multilingual setting.

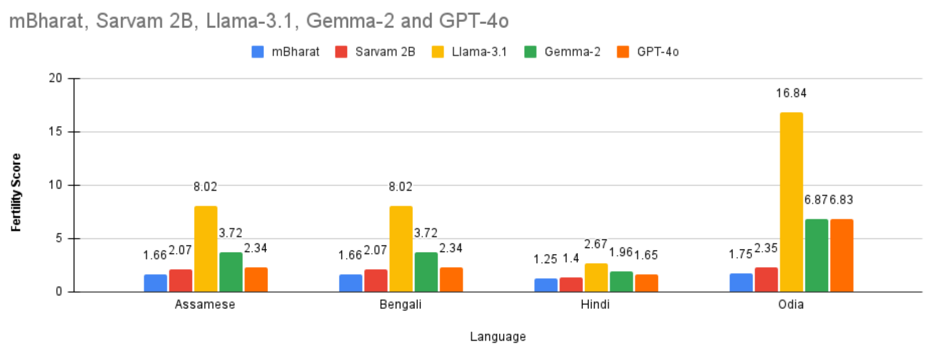


Figure 2: Fertility score of mBharat tokenizer v/s LLMs across languages of 3 scripts (Bengali, Devanagari, and Odia). LLMs score are reported from (Sarvam2B, 2024)

2019) with modifications. The model architecture uses RMSNorm as pre-normalizaion layer with norm_epsilon = 1e-5, approximate GeGLU Activations (Shazeer, 2020) as non-linearity by replacing the standard ReLU non-linearity activation function. The model architecture uses a scaled version of RoPE embedding (Su et al., 2022) with θ=10,000. We scaled the RoPE embedding through a shrinking factor by dividing the target context length y by $permissible\_context\_size\_length$ on single GPU, keeping all other hyperparameters fixed such as batch size, vocabulary size, etc. This allows every $position\_ids$ to be divided by the shrinking ratio in the RoPE embedding methodology. For instance, if the $permissible\_context\_size\_length$ for a given physical memory hardware is 256, then we apply shrinking factor of 16 for target context size of 4096 on Single A100 40G chip during pretraining. Then, a token with $position\_ids$ = 4000 becomes 4000/16 = 250, and the neighbouring token 4001 becomes 4001/16 = 250.06, to be within 0 to 256. This is how we can capture higher context size during pretraining on limited physical memory required to pretrain model at higher context size outside the $permissible\_context\_size\_length$. This modification allows us to pretrain models from scratch at much higher context size than the physical memory required for pretraining. Hence, with limited physical memory and limited GPUs, we can pretrain generative language models from scratch at much higher desired context size. Following (Chowdhery et al., 2023), we remove all biases from dense layers to improve the training stability. Table 17 lists down the various configuration to develop our 13.29M, 26.59M, 87M, 87.25M, 92.63M, 108.5M, 139.3M, 162M, 208M, 350M models. Table 17 lists the different model configuration but we used weight tying (Press & Wolf, 2017) to improve the performance of language models by sharing the weights of the embedding and softmax layers. Therefore, the total number of parameters in our models are typically smaller due to weight tying and different tokenizer sizes of each language specific model than the sizes mentioned in Table 17.

3.5 Training

We performed hyperparameter tuning on 15M models to find the optimal tokenizer size for each language specific monolingual corpus, learning rate, learning rate scheduler, and weight decay. We used the concept of μP transfer (Yang et al., 2021), and transferred the learned hyperparameters to our bigger models. For further training and hyperparameter tuning details, we mentioned in

Appendix A.3. We find the optimal size of tokenizers by training range of tokenizers of various sizes and optimizing the validation perplexity of the language models and fertility score of the tokenizers.

3.6 Impact of language specific merged-tokenizer v/s language agnostic tokenizer for Bilingual model (Konkani-Maithili 27M GPT) of typologically same script (Devanagari)

For the language agnostic approach, we merged both the Konkani and Maithili corpus which are of comparable size after our preprocessing in terms of number of sentences (lines) then we trained SentencePiece's BPE algorithm to a vocabulary size of 1750 on the whole corpus and trained our 27M Konkani-Maithili bilingual GPT model and we report the validation perplexity score to be 12.43393. For the language specific approach, we trained individual SentencePiece (Kudo & Richardson, 2018) BPE tokenizers independently on Konkani (vocabulary size of 1k) and Maithili (vocabulary size of 750) corpus respectively. After that, we merged both the language specific tokenizers and also removing the overlapped tokens as both Konkani and Maithili are written in Devanagari script. We used this merged tokenizer to train another 27M bilingual Maithili-Konkani GPT and found the validation perplexity of the bilingual modeling drops to 8.53827 from 12.43393. From our experiments, we infer that language specific tokenization is very important to preserve the language specific linguistics even for thre typologically similar languages (Devanagari script) and the merging operation of the tokenization also helps to omit the overlapping tokens which results in optimized vocabulary size for language modeling resulting in better open-end text generation, lower perplexity score due to optimal size of the tokenizer for multilingual language modeling.

3.7 Language Modeling for Paramanu-Konkani (Devanagari script) 15M with BPE tokenizer v/s Unigram tokenizer

From our experiments where we tokenized the Konkani corpus with Sentencepiece's Unigram model with a vocabulary size of 1000 and trained a 15 Model and similarly we also trained another 15M model with SentencePiece BPE tokenizer on the same setting to see whether there is an impact of these two different tokenizers on the perplexity score of monolingual language modeling. We found the validation perplexity for Unigram 15M model to be 11.88412 whereas for BPE 15M model is 11.74314 which is slightly lower but not significantly different. However, Unigram tokenizer takes longer time than BPE to get trained on the corpus.

3.8 Quantitative Results based on Perplexity Metric

Table 2 lists the validation perplexity and MFU metrics of our various pretrained models. In terms of quantitative evaluation of language modeling, the lower the perplexity, the better is the language model. From Table 2, Paramanu-Sanskrit exhibited the lowest validation perplexity among our models, followed by Odia, Bangla, Telugu, mParamanu (162M), Tamil, Marathi, and Hindi, which had the highest perplexity. It is important to note that most models were trained for the same number of steps, regardless of their pretraining dataset size, meaning some may be under-trained. Continued pretraining, such as for Paramanu-Hindi, could further improve perplexity scores.

3.9 Based on Quantitative Benchmark

We evaluated our models on key Natural Language Understanding (NLU), Natural Language Inference, and Commonsense Reasoning tasks, including their multilingual variants, while benchmarking against multilingual LLMs like Sarvam-2B, Bloom, and Bloomz in two groups, one with LLMs of size <=2B and another group of LLMs of size > 2B. We utilized the translated ARC, HellaSwag, and MMLU datasets from (Lai et al., 2023), employing Eleuther AI's LM Evaluation Harness (Sutawika et al., 2024) for evaluation. Key assessments included HellaSwag (Zellers et al., 2019), which tests common sense reasoning by predicting scenario endings; MMLU (Hendrycks et al., 2021), measuring broad knowledge across diverse subjects; and ARC-Challenge (Clark et al., 2018), which examines complex reasoning with scientific questions. We encountered misalignment issues with the LM Evaluation Harness datasets and our models, preventing evaluations on HellaSwag except for Hindi. Other evaluations included XCOPA (Ponti et al., 2020), assessing cross-lingual commonsense reasoning; XNLI (Conneau et al., 2018), which assesses cross-lingual sentence classification

Model	Perplexity
Paramanu-Assamese 26.59M	6.620
Paramanu-Bangla 87.25M	5.069
Paramanu-Bangla 108.5M	4.102
Paramanu-Hindi 162M	16.992
Paramanu-Hindi 367.5M	11.052
Paramanu-Konkani-Maithili 13.29M (merged language specific tokenizer)	8.538
Paramanu-Konkani-Maithili 13.29M (language agnostic tokenizer)	12.433
Paramanu-Odia 87M	3.068
Paramanu-Sanskrit 139.33M	1.748
mParamanu 92.63M	8.443
mParamanu 162M	6.924
Paramanu-Marathi 207.73M	8.943
Paramanu-Telugu 208.25M	5.400
Paramanu-Tamil 207.84M	7.618

Table 2: Perplexity of models

N-shot	XNLI-Hindi	XStoryCloze-Hindi	XStoryCloze-Telugu	XCOPA-Tamil
0	33.49	**52.42**	**56.06**	**54.00**
5	**34.04**	51.49	54.67	52.40
25	33.23	52.02	55.92	49.80

Table 3: N-shot evaluation of pretrained Paramanu models across various benchmarks.

across 15 languages; and XStoryCloze (Lin et al., 2022), evaluating story understanding by selecting the correct ending to a four-sentence story. Together, these benchmarks comprehensively assess model performance and reasoning capabilities.

Table 4, Table 5, Table 6, Table 7 and Table 8 evaluate model performance in a zero-shot setting using accuracy metrics across translated benchmarks (ARC, MMLU, HellaSwag) for Bangla, Hindi, Tamil, and Telugu. Table 3 presents n-shot evaluations for XNLI in Hindi, XStoryCloze in Hindi and Telugu, and XCOPA in Tamil. Additionally, Table 5 and Table 6 assess various pretrained monolingual and multilingual models for Devanagari script across MMLU, HellaSwag, ARC, XStoryCloze, XNLI, and Belebele, highlighting cross-lingual language transfer among Devanagari languages (Hindi, Marathi). Notably, mParamanu, pretrained on low-resource Devanagari languages (Sanskrit, Konkani, Maithili), achieved scores of 25.86 for MMLU-Marathi, 24.84 for Hindi, 28 for Belebele-Marathi, and 25.44 for Belebele-Hindi, indicating effective language transfer to medium (Marathi) and high-resource (Hindi) languages using the mBharat tokenizer. Interestingly, neither Paramanu-Sanskrit nor mParamanu were pretrained on Hindi or Marathi but still performed well on their benchmarks. This is possibly due to the same script.

Table 5, Table 13 in Appendix and Table 6, Table 14 in Appendix show that Paramanu models exhibit superior performance across various benchmarks in Devanagari languages, despite their smaller sizes and being pretrained on fewer tokens than larger multilingual LLMs. Specifically, Paramanu-Marathi (208M) outperformed Sarvam (2B), OpenHathi (7B), and Bloom (560M) on the Marathi benchmark. Similarly, mParamanu (162M) outperformed Paramanu-Hindi (367M) and demonstrated competitiveness against larger models. Notably, Paramanu-Hindi-instruct (356M) surpassed all larger multilingual LLMs, except Bloomz (7B), by a significant margin, benefiting from instruction tuning on a dataset of 27,000 Hindi instructions and additional 52,000 Alpaca machine-translated instructions. In contrast, Bloomz (7B) was trained on hundreds of thousands of instructions. If the 367M Hindi model underwent more training steps, it could potentially achieve even better performance, as many models were pretrained for the same duration regardless of dataset size. The stronger performance of mParamanu in Hindi illustrates effective language transfer within the same script and typology. Notably, Paramanu-Sanskrit (139M), pretrained on 45 billion tokens, achieved an average score of 31.05, surpassing both Hindi models and closely approaching Bloom (560M) and Bloomz (560M). Its lower perplexity (1.75) compared to Paramanu-Hindi (11.05) further supports the notion that additional pretraining for the Hindi model could significantly enhance its downstream performance.

From Table 3, we observe that the performance of our models drop from zero-shot setting to 25 shot setting on XNLI-Hindi, XStoryCloze for Hindi and Telugu, and XCOPA for Tamil. This type of phenomenon has also been observed in PlanningBench (Valmeekam et al., 2024) where GPT-3.5-Turbo, GPT-4, GPT4-o performance on Blocksworld dropped from 0 shot to 1-shot significantly.

Models	MMLU-Bangla	ARC-Bangla	Belebele-Bangla	Average (Bangla)	Belebele-Assamese
Paramanu-Bangla 108M (ours)	23.82	25.75	25.11	24.89	25.33
Paramanu-Bangla-instruct 108M (ours)	**27.60**	**28.50**	**32.45**	**29.52**	**30.54**
mParamanu 162M (ours)	25.29	20.19	27.44	24.31	29.00
Bloom 560M	22.61	26.00	22.89	23.83	22.78
Bloomz 560M (instruction-tuned)	25.82	23.43	22.77	24.01	25.11
Bloom 1.1B	23.90	24.37	26.00	24.75	26.89
Sarvam 2B	24.05	28.40	23.22	25.22	27.78

Table 4: Zero-shot evaluation of LLMs (<=2B) across translated benchmarks of MMLU, HellaSwag, ARC datasets, and Belebele in Bengali script. All benchmarks report Accuracy except for ARC which reports Normalized Accuracy. Max scores are in bold.

Models	MMLU-Marathi	ARC-Marathi	Belebele-Marathi	Average (Marathi)
mParamanu 162M (ours)	25.68	22.16	28.00	25.28
Paramanu-Hindi 367M (ours)	23.78	24.16	24.66	24.20
Paramanu-Hindi-instruct 367M (ours)	**28.72**	**27.85**	**32.00**	**29.52**
Paramanu-Marathi 208M (ours)	25.39	26.49	27.33	26.40
Paramanu-Sanskrit 139M (ours)	24.96	26.49	24.33	25.26
Bloom 560M	22.78	24.50	27.00	24.76
Bloomz 560M (instruction-tuned)	26.20	24.24	25.44	25.29
Bloom 1B	23.93	25.10	28.33	25.78
Sarvam 2B	23.96	27.53	26.77	26.08

Table 5: Zero-shot evaluation of LLMs (<=2B) for cross-lingual language transfer in Marathi. All benchmarks report Accuracy except for ARC (Normalized Accuracy). Max scores are in bold.

Perhaps n-shot examples become additional soft constraints on the generation which might be the reason of degradation of performance from the original training dataset. From Table 9, we see that Paramanu-Bangla 108M outperformed Bloom 560M by 1.06% points, Bloomz 560M by 1.05% points, Bloom 1.1B on average score across MMLU, ARC, and Belebele benchmarks, and by 1.21% points on MMLU over Bloom 560M depsite being smaller by 10.2 times compared to Bloom 1.1B and being pretrained on 26.21 billon of tokens. However, Paramanu-Bangla 108M is extensively trained only on Bangla literature corpus. With further instruction-tuning on 27k Bangla instructions, Paramanu-Bangla-instruct 108M outperformed Bloom 560M, Bloomz 560M, Bloom 1.1B, Sarvam 2B, Bloom 7B on average score of MMLU, ARC, and Belebele benchmarks for Bangla respectively.

Table 7 and Table 8 compare pretrained multilingual LLMs (<2B) and instruction-tuned models on Tamil and Telugu benchmarks. Table 15 and Table 16 in Appendix compares our model with LLMs (> 2B) Our model, Paramanu-Tamil (208M), outperformed larger multilingual LLMs like Bloom (560M), Bloomz (560M), and Bloom (1.1B) across four benchmarks (Belebele, XCOPA, MMLU, and ARC) in Tamil, coming close to Sarvam (2B) despite being much smaller and trained on fewer tokens. On MMLU-Tamil, both Paramanu-Tamil and Paramanu-Tamil-instruct outperformed Sarvam (2B) by 2.89 percentage points, with Paramanu-Tamil pretrained on 26.2 billion tokens. Paramanu-Tamil-instruct surpassed Bloom (7B), despite being 34 times smaller, and outperformed Bloom (1.1B) by 2.28 points, Bloom (560M) by 3 points, and Bloomz (560M) by 1.17 points. For Telugu, Paramanu-Telugu-instruct (208M) outperformed Bloom (560M) by 2.72 points, Bloomz by 2.25 points, Bloom (1.1B) by 1.46 points, and Bloom (7B) by 1.22 points, with Paramanu-Telugu pretrained on 39.32 billion tokens. The improvements in metric scores for Tamil and Telugu instruction-tuned models were modest, likely due to lower-quality machine translations from Bangla compared to Hindi. Nonetheless, these results show strong performance of our models on various NLP tasks despite their smaller size and fewer tokens, challenging the notion that larger models are always better. Our findings suggest that smaller pretrained models can excel when trained on high-quality, preprocessed data over multiple epochs, outperforming larger models trained on lower-quality data for an epoch.

3.9.1 Based on Human Evaluation

We hard-prompted various LLMs (GPT-2 XL, GPT Neo 1.3B, LLaMa 2 7B, OPT 6.7B, and the multilingual Bloom series) alongside our pretrained models (Paramanu-Bangla, Paramanu-Hindi, and mParamanu for Sanskrit) without fine-tuning. The prompts reflected the local, cultural, and literary contexts of Assamese, Bangla, Hindi, Konkani, Maithili, Odia, and Sanskrit. Due to resource constraints, extensive evaluations focused on Paramanu-Bangla, Paramanu-Hindi, and mParamanu. The top three predictions from each model were generated with temperature $= 1.0$ and $\text{top}_p = 0.9$.

Models	MMLU-Hindi	HellaSwag-Hindi	ARC-Hindi	XStoryCloze-Hindi	XNLI-Hindi	Belebele-Hindi	Average (Hindi)
mParamanu 162M (ours)	24.84	24.87	22.35	49.24	33.70	25.44	30.07
Paramanu-Hindi 367M (ours)	24.38	24.83	27.05	47.92	32.00	23.33	29.92
Paramanu-Hindi-instruct 367M (ours)	**30.25**	29.42	**30.23**	58.00	40.25	**42.78**	**40.14**
Paramanu-Marathi 208M (ours)	25.49	26.59	23.97	48.71	33.73	27.33	30.97
Paramanu-Sanskrit 139M (ours)	25.16	25.64	25.17	50.23	34.46	25.66	31.05
Bloom 560M	23.67	27.50	23.88	54.79	40.80	26.44	32.84
Bloomz 560M (instruction-tuned)	25.87	26.48	24.40	55.53	35.58	26.00	32.31
Bloom 1B	23.86	28.28	24.74	55.59	42.77	28.00	33.87
Sarvam 2B	24.54	**33.66**	28.00	**60.29**	**46.74**	24.44	36.27

Table 6: Zero-shot evaluation of LLMs(<=2B) for cross-lingual language transfer in Hindi. All benchmarks report Accuracy except for ARC (Normalized Accuracy). Max scores are in bold.

Models	Belebele-Tamil	XCOPA-Tamil	MMLU-Tamil	ARC-Tamil	Average (Tamil)
Paramanu-Tamil 208M (ours)	26.88	57.60	24.37	24.51	33.34
Paramanu-Tamil-instruct 208M (ours)	**30.22**	56.00	**26.95**	26.04	34.80
Bloom 560M	27.22	55.80	23.95	25.57	33.13
Bloomz 560M (instruction-tuned)	23.55	58.60	25.78	25.30	33.30
Bloom 1.1B	25.77	57.00	24.67	24.34	32.94
Sarvam 2B	27.44	**63.00**	24.06	**26.53**	**35.25**

Table 7: Zero-shot evaluation of LLMs (<=2B) in Tamil script models. All benchmarks report Accuracy except for ARC which reports Normalized Accuracy. Max scores are in bold.

~~Native speakers assessed the outputs on Grammar, Coherence, Creativity, and Factuality, each scored from 0 (worst) to 5 (best).~~ For human evaluation, we asked 10 annotators to evaluate top-3 responses from models for each prompt on a scale of 0 (worst) to 5 (best). We report the average score of all ratings. We also have reported normalised scores of ratings in Table 19 in appendix to handle inconsistencies among annotators. We reached inter-annotator kappa score of 0.85 for Bengali, 0.79 for Hindi, and 0.72 for Sanskrit. Figure 4 in the Appendix shows the bar chart for inter-annotator agreement's Fleiss Kappa score. For Factuality, higher scores indicated better alignment with real events, with some evaluators assigning a score of 0 when premises could not be verified.

Table 24 displays the human evaluation of Paramanu-Bangla 87.25M model for the mentioned Bangla prompts. Paramanu-Bangla 87.25M model scored an average score of 3.5/5 on grammar, 3.325/5 on coherence, 3.225/5 on creativity, and 3.2/5 on factuality metrics across top 3 generations for each Bangla prompt. Table 10 compares the performance of Paramanu-Hindi 162M and other LLMs including multilingual Bloom which was pretrained on Indian languages. In this table, we can complete see that none of the open source LLMs have the ability to generate grammatically, coherent sentences in Hindi except the Bloom series. Our monolingual model, Paramanu-Hindi 162M has performed better by 17.25% on grammar, by 46.05% on coherence, by 62.5% on creativity, and by 238.5% on factuality compared to Bloom 3B model despite being 19 times smaller in size. Table 11 in Appendix compares the performance of mParamanu-162M and other LLMs including multilingual Bloom which was pretrained on Indian languages. We can see that none of the LLMs have the ability to generate grammatically, coherent sentences in Sanskrit keeping aside the factuality. Our multilingual model, mParamanu-162M has scored the highest among all on grammar (3.75/5), coherence (3.166/5), creativity (2.166/5), and factuality (1.75/5) whereas Bloom 3B scored 0.166/5 on grammar, 0.0833/5 on coherence, and 0/5 for both creativity and factuality metrics respectively. GPT-3.5-Turbo (ChatGPT) has scored very poorly 0.25/5 on grammar & coherence metrics, 0.1818/5 in creativity and 0.33/5 on factuality metrics respectively for Sanskrit text generation. mParamanu-162M is smaller by 44.25 times compared to 7B LLaMa-2 model and yet it has shown its high quality text generation in Sanskrit as compared to ChatGPT, LLaMa, and Bloom series of models.

4 Conclusions and Future Work

In this paper, we present a series of novel efficient small monolingual, bilingual, and multilingual pretrained auto regressive models the range of 13.5M to 367.5M for 10 Indian languages across 5 scripts excluding English, thereby enabling GenAI NLP technology accessible to billion speakers in the world and working towards including underrepresented languages to reduce the language bias and divide in today's NLP technology. We proposed a RoPE embedding scaling method that enables us to pretrain language models from scratch at larger sequence length context without increased GPU memory. We proposed novel tokenization technique of combining both BPE and Unigram tokenizers into a single tokenizer. We also proposed and performed language specific tokenization for

Models	Belebele-Telugu	XStoryCloze-Telugu	MMLU-Telugu	ARC-Telugu	Average (Telugu)
Paramanu-Telugu 208M (ours)	26.00	51.42	25.12	**26.32**	32.22
Paramanu-Telugu-instruct 208M (ours)	**27.50**	58.00	**26.75**	25.75	**34.50**
Bloom 560M	23.55	55.65	24.10	23.85	31.78
Bloomz 560M (instruction-tuned)	22.44	54.86	26.82	24.91	32.25
Bloom 1.1B	26.88	56.38	24.53	24.38	33.04
Sarvam 2B	**27.66**	**60.09**	24.67	25.78	**34.55**

Table 8: Zero-shot evaluation of LLMs (<=2B) in Telugu script models. All benchmarks report Accuracy except for ARC which reports Normalized Accuracy. Max scores are in bold.

Model	Grammar	Coherence	Creativity	Factuality
GPT2-XL	0.45833	0.37500	0.37500	0.37500
GPT-Neo 1.3B	0.91666	0.91666	0.91666	0.91666
OPT 6.7B	0.70833	0.70833	0.70833	0.70833
GPT-J 6B	1.12500	0.95833	0.95833	0.95833
LLaMa 2 7B	0.70833	0.70833	0.70833	0.70833
Bloom 560M	1.70833	1.41666	1.41666	1.37500
Bloom 1.1B	1.33333	1.29166	1.29166	1.29166
Bloom 3B	1.54166	1.29166	1.33333	1.33333
Bloom 7.1B	1.75000	1.16666	1.16666	1.08333
GPT-3.5-Turbo	0.58330	0.75000	0.58330	0.58330
Paramanu-Bangla 108.5M (ours)	**4.66666**	**4.58333**	**3.76280**	**3.45833**

Table 9: Human evaluation results (average scores of top-3 generations per prompt) of open-end text generation of Paramanu-Bangla v/s LLMs for 4 Bangla prompts on Grammar, Coherency, Creativity, and Factuality metrics. Scale is 0 (worst) to 5 (best). GPT-3.5-Turbo was accessed in October 2023.

multilingual language modeling and domain specific tokenization for monolingual language modeling for strong performance. Our multilingual tokenizer, mBharat shows the best fertility scores among Indian language tokenizers. We considered typological grouping and pretraining on comparable size of each monolingual language corpus for our multilingual mParamanu 162M to handle data imbalance and curse of multilinguality. We evaluated our models for open-end text generation with human evaluators on grammar, coherency, creativity, and factual metrics. We reached inter-annotator kappa score of 0.85 for Bengali, 0.79 for Hindi, and 0.72 for Sanskrit. In our evaluation, we found that none of the popular existing LLMs can generate grammatically correct and coherent sentences in 10 Indian languages despite being pretrained on Indian language corpora. Our efficient generative language models have performed better than Bloom 7B, LLaMa-2 7B, OPT 6.7B, GPT-J 6B, GPTNeo 1.3B, GPT2-XL models for open-end text generation in Assamese, Bangla, Hindi, Odia, and Sanskrit despite being 66 times to 20 times smaller in size. We also evaluated our models and compared with several multilingual LLMs across various NLU, NLI, and commonsense reasoning benchmarks. Our models outperformed most multilingual LLMs of size 2B and performed very competitive or even better than LLMs of size 7B on various LLM benchmarks despite being smaller in size by by multiple order of magnitude ~~compared with LLMs whose size is bigger by multiple order of magnitude.~~ We observed language transfer phenomena from low-resource to high resource languages of same script and typology. We also instruction-tuned our pretrained models for Bangla, Hindi, Marathi, Tamil, and Telugu and show their task handling capabilities.

In future, we would like to extend our multilingual model to 22 official Indian languages and align our generative language models with multimodal encoders to develop multimodal generative language models for Indian languages.

Model	Grammar	Coherence	Creativity	Factuality
GPT2-XL	0.0000	0.0000	0.0000	0.0000
GPT-Neo 1.3B	0.0000	0.0000	0.0000	0.0000
OPT 6.7B	0.5833	0.1667	0.1667	0.0000
GPT-J 6B	0.5833	0.2500	0.0000	0.0000
LLaMa 2 7B	1.3333	0.3333	0.5000	0.2083
Bloom 560M	2.7917	2.4583	1.0000	1.1667
Bloom 1.1B	3.2917	2.7917	1.6250	1.3333
Bloom 3B	4.0833	3.1666	2.0000	1.0833
Bloom 7.1B	3.2917	2.7917	1.6250	1.3333
Paramanu-Hindi 162M (ours)	**4.7917**	**4.6250**	**3.2500**	**3.6666**

Table 10: Human Evaluation results (avg scores of top 3 generations per prompt) of open-end text generation of Paramanu-Hindi 162M v/s LLMs for 4 Hindi prompts on Grammar, Coherency, Creativity, and Factuality metrics. Scale is 0 (worst) to 5 (best)

REFERENCES

Google Translate Cloud REST API. https://cloud.google.com/translate/docs/reference/rest, 2023. Accessed: 2023-10-30.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4762–4781, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.410. URL https://aclanthology.org/2021.findings-emnlp.410.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

Pramit Bhattacharyya, Joydeep Mondal, Subhadip Maji, and Arnab Bhattacharya. VACASPATI: A diverse corpus of Bangla literature. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1118–1130, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.72. URL https://aclanthology.org/2023.ijcnlp-main.72.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=OUIFPHEgJU.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12402–12426, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.693. URL https://aclanthology.org/2023.acl-long.693.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha

Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Aswanth Kumar M, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M. Khapra, Raj Dabre, Rudra Murthy, and Anoop Kunchukuttan. Airavata: Introducing hindi instruction-tuned llm, 2024. URL https://arxiv.org/abs/2401.15006.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference* *on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022b.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates, 2018.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback, 2023.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4483–4499, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.363. URL https://aclanthology.org/2020.emnlp-main.363.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616.

Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, Yanqi Zhou, Wei Li, Nan Ding, Jake Marcus, Adam Roberts, and Colin Raffel. Do transformer modifications transfer across implementations and applications? In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5758–5773, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.465. URL https://aclanthology.org/2021.emnlp-main.465.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In Marine Carpuat, Marie-Catherine de Marneffe, and

Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1791–1799, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.130. URL https://aclanthology.org/2022.naacl-main.130.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185.

Ofir Press and Lior Wolf. Using the output embedding to improve language models, 2017.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

Sarvam2B. sarvamai/sarvam-2b-v0.5 ů Hugging Face — huggingface.co. https://huggingface.co/sarvamai/sarvam-2b-v0.5, 2024. [Accessed 15-09-2024].

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Noam Shazeer. Glu variants improve transformer, 2020.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2022.

Lintang Sutawika, Hailey Schoelkopf, Leo Gao, Stella Biderman, Baber Abbasi, Jonathan Tow, ben fattori, Charles Lovering, farzanehnakhaee70, Jason Phang, Anish Thite, Fazz, Aflah, Niklas Muennighoff, Thomas Wang, sdtblck, gakada, nopperl, researcher2, tttyuntian, Chris, Julen Etxaniz, Zdenk Kasner, Khalid, Jeffrey Hsu, Hanwool Albert Lee, Anjor Kanekar, AndyZwei, Pawan Sasanka Ammanamanchi, and Dirk Groeneveld. Eleutherai/lm-evaluation-harness: v0.4.1, January 2024. URL https://doi.org/10.5281/zenodo.10600400.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogoziska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluciska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: an extensible benchmark for evaluating large language models on planning and reasoning about change. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 60006010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. Neurons in large language models: Dead, n-gram, positional. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1288–1301, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.75. URL https://aclanthology.org/2024.findings-acl.75.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2023.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario ako, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL https://arxiv.org/abs/2211.05100.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Bx6qKuBM2AD.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:159041722.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

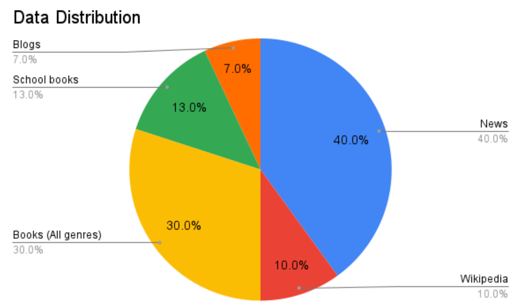


Figure 3: Pretraining Data distribution.

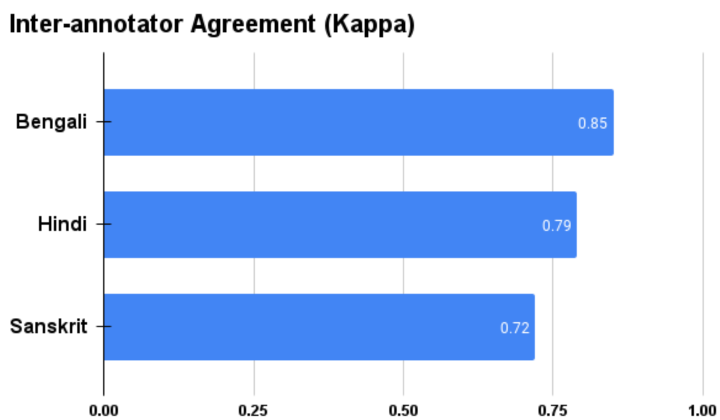


Figure 4: Inter-Annotator Agreement (Kappa)

A Appendix

Model	Grammar	Coherence	Creativity	Factuality
GPT2-XL	0.17	0.08	0.00	0.00
GPT-Neo 1.3B	0.25	0.00	0.00	0.00
OPT 6.7B	0.00	0.00	0.00	0.00
GPT-J 6B	0.33	0.33	0.00	0.00
LLaMa 2 7B	0.42	0.33	0.42	0.00
Bloom 560M	0.08	0.00	0.08	0.00
Bloom 1.1B	0.08	0.00	0.00	0.00
Bloom 3B	0.17	0.08	0.00	0.00
GPT-3.5-Turbo	0.25	0.25	0.18	0.33
mParamanu 162M (ours)	**3.75**	**3.17**	**2.17**	**1.75**

Table 11: Human evaluation results (average scores for top-3 generations per prompt) of open-end text generation of mParamanu v/s LLMs for 4 Sanskrit prompts on various metrics. Scale is from 0 (worst) to 5 (best). GPT-3.5-Turbo was accessed in October 2023.

A.1 Model parameters

A.2 Language Demographics

A.3 Training

Following (Hoffmann et al., 2022b), we set lr decay steps to max_steps and the minimum lr is set nearly to $0.1{\cdot}lr$. The lr schedule starts with a linear warm-up from 0 to the maximum lr at 1000

Models	MMLU-Bangla	ARC-Bangla	Belebele-Bangla	Average (Bangla)	Belebele-Assamese
Paramanu-Bangla 108M (ours)	23.82	25.75	25.11	24.89	25.33
Paramanu-Bangla-instruct 108M (ours)	27.60	28.50	32.45	29.52	30.54
mParamanu 162M (ours)	25.29	20.19	27.44	24.31	29.00
Bloom 7B	27.10	26.09	23.22	25.47	23.11
Bloomz 7B (instruction-tuned)	32.46	27.20	53.67	37.77	48.00

Table 12: Zero-shot evaluation of LLMs (> 2B) across translated benchmarks of MMLU, HellaSwag, ARC datasets, and Belebele in Bengali script. All benchmarks report Accuracy except for ARC which reports Normalized Accuracy.

Models	MMLU-Marathi	ARC-Marathi	Belebele-Marathi	Average (Marathi)
mParamanu 162M (ours)	25.68	22.16	28.00	25.28
Paramanu-Hindi 367M (ours)	23.78	24.16	24.66	24.20
Paramanu-Hindi-instruct 367M (ours)	**28.72**	**27.85**	32.00	29.52
Paramanu-Marathi 208M (ours)	25.39	26.49	27.33	26.40
Paramanu-Sanskrit 139M (ours)	24.96	26.49	24.33	25.26
Bloom 7B	27.30	25.54	24.00	25.61
Bloomz 7B (instruction-tuned)	32.62	27.44	53.00	37.68
OpenHathi 7B	26.09	24.24	25.88	25.40
Airavata 7B (instruction-tuned)	26.15	23.90	29.89	26.64

Table 13: Zero-shot evaluation of LLMs (>2B) for cross-lingual language transfer in Marathi. All benchmarks report Accuracy except for ARC which reports Normalized Accuracy.

steps, followed by a cosine decay to the minimum lr until the end of an epoch of training. We used the following equation for lr decay ratio.

$$lr_{decay\_ratio} = \frac{t - warmup_{steps}}{lr_{decay\_steps} - warmup_{steps}}$$

where t is the current training step.

A.3.1 26.58M PARAMANU-ASSAMESE MODEL

We used the same training procedure as mentioned in 3.5 but with a batch size of 64, gradient accumulation steps of 4, and the maximum sequence length set to 1024, i.e., 262,144 tokens per iteration and transferred the learned hyperparameters from 15M model to 42M model using (μP) transfer. We set maximum learning rate (lr) to 3e-3 (max), weight decay to 1e-1. We trained our bigger models with fused $AdamW$ optimizer for an epoch of training with β_1=0.9, β_2=0.95, dropout of 0.0, and scaled the gradient norms using a maximum norm clipping value of 1.0, and weight decay of 0.1. For our experiments and modeling, we implemented our code using Pytorch 2.0, in-house optmized CUDA kernels and used `torch.compile` feature for every model. To further speedup training, we used BF16 mixed precision training.

A.3.2 87.25M PARAMANU-BANGLA MODEL

We used the same training procedure as mentioned in 3.5 but with a batch size of 32, gradient accumulation steps of 8, and the maximum sequence length set to 1024, i.e., 262,144 tokens per iteration and transferred the learned hyperparameters from 15M model to 110M model using (μP) transfer.

A.3.3 108.5M PARAMANU-BANGLA MODEL

We used the same training procedure as mentioned in 3.5 but with a batch size of 32, gradient accumulation steps of 8, and the maximum sequence length set to 1024, i.e., 262,144 tokens per iteration and transferred the learned hyperparameters from 15M model to 140M model using (μP) transfer.

A.3.4 162M PARAMANU-HINDI MODEL

We used the same training procedure as mentioned in 3.5 but with a batch size of 32, gradient accumulation steps of 8, the maximum sequence length set to 1024, i.e., 262,144 tokens per iteration and transferred the learned hyperparameters from 15M model to 162M model using (μP) transfer.

Models	MMLU-Hindi	HellaSwag-Hindi	ARC-Hindi	XStoryCloze-Hindi	XNLI-Hindi	Belebele-Hindi	Average (Hindi)
mParamanu 162M (ours)	24.84	24.87	22.35	49.24	33.70	25.44	30.07
Paramanu-Hindi 367M (ours)	24.38	24.83	27.05	47.92	32.00	23.33	29.92
Paramanu-Hindi-instruct 367M (ours)	30.25	29.42	30.23	58.00	40.25	42.78	40.14
Paramanu-Marathi 208M (ours)	25.49	26.59	23.97	48.71	33.73	27.33	30.97
Paramanu-Sanskrit 139M (ours)	25.16	25.64	25.17	50.23	34.46	25.66	31.05
Bloom 7B	27.04	31.39	26.36	60.55	47.18	23.00	35.92
Bloomz 7B (instruction-tuned)	35.55	28.57	29.36	57.71	40.52	53.11	40.80
OpenHathi 7B	27.69	30.54	25.51	57.04	39.03	32.66	35.41
Airavata 7B (instruction-tuned)	30.43	29.53	25.60	55.59	39.04	41.44	36.93

Table 14: Zero-shot evaluation of LLMs (>2B) for cross-lingual language transfer in Hindi. All benchmarks report Accuracy except for ARC which reports Normalized Accuracy.

Models	Belebele-Tamil	XCOPA-Tamil	MMLU-Tamil	ARC-Tamil	Average (Tamil)
Paramanu-Tamil 208M (ours)	26.88	57.60	24.37	24.51	33.34
Paramanu-Tamil-instruct 208M (ours)	30.22	56.00	26.95	26.04	34.80
Bloom 7B	25.55	59.20	26.39	24.69	33.95
Bloomz 7B (instruction-tuned)	50.66	57.40	29.48	28.10	41.41

Table 15: Zero-shot evaluation of LLMs (>2B) in Tamil script models. All benchmarks report Accuracy except for ARC which reports Normalized Accuracy.

A.3.5 367.5M Paramanu-Hindi model

We used the same training procedure as mentioned in 3.5 but with a batch size of 32, gradient accumulation steps of 16, the maximum sequence length set to 1024, i.e., 524,288 tokens per iteration and transferred the learned hyperparameters from 15M model to 367.5M model using (μP) transfer. After 1 epoch of training, the average validation perplexity is 11.05240 whereas the average training perplexity is 10.99616.

A.3.6 87M Paramanu-Odia model

We used the same training procedure as mentioned in 3.5 but with a batch size of 64, gradient accumulation steps of 8, and the maximum sequence length set to 1024, i.e., 524,288 tokens per iteration and transferred the learned hyperparameters from 15M model to 110M model using (μP) transfer.

A.3.7 139.3M Paramanu-Sanskrit model

We used the same training procedure as mentioned in 3.5 but with a batch size of 64, gradient accumulation steps of 8, and the maximum sequence length set to 1024, i.e, 524,288 tokens per iteration and transferred the learned hyperparameters from 15M model to 175M model using (μP) transfer.

A.3.8 13.5M bilingual Konkani-Maithili GPT model

Both Konkani and Maithili languages are typologically similar (Devanagari script). For bilingual model, we trained two 27M models with language agnostic tokenization and language specific tokenization to study the impact of language specific tokenization against language agnostic tokenization for multilingual language modeling. For training with the language specific tokenization, where we basically trained independent BPE tokenizer on Konkani and Maithili corpora with a tokenizer size of 1000 and 750 respectively and merged them together. For language agnostic tokenization, we trained BPE tokenizer on the merged corpora of Konkani and Maithili with a tokenizer size of 1750.

We used the same training procedure as mentioned in 3.5 but with a batch size of 128, gradient accumulation steps of 2, and the maximum sequence length set to 1024, i.e., 262,144 tokens per iteration . We transferred the learned hyperparameters from 15M model to 27M model using (μP) transfer.

A.3.9 92.63M multilingual **mParamanu model**

For our multilingual mParamanu model, we selected the languages on the basis of typological grouping and having comparable corpora as shown in the Table 19, to avoid pretraining our multilingual model on disproportionate corpora of multiple languages. Thus, we avoid adding the Hindi corpus

Models	Belebele-Telugu	XStoryCloze-Telugu	MMLU-Telugu	ARC-Telugu	Average (Telugu)
Paramanu-Telugu 208M (ours)	26.00	51.42	25.12	26.32	32.22
Paramanu-Telugu-instruct 208M (ours)	27.50	58.00	26.75	25.75	34.50
Bloom 7B	24.66	57.37	26.62	24.47	33.28
Bloomz 7B (instruction-tuned)	43.11	58.23	29.55	27.98	39.71

Table 16: Zero-shot evaluation of LLMs (> 2B) in Telugu script models. All benchmarks report Accuracy except for ARC which reports Normalized Accuracy.

Test input indic multilingual text in 7 languages: ভাৰত মোৰ মাতৃভূমি ভারত আমার মাতৃভূমি साडा देश भारत है भारत हो म्हजो देश भारत हमर देश अछि ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି अस्माकं देशः भारतवर्षः

Tokenized by LLaMA-2 32k tokenizer:['_', '<0xE0>', '<0xA6>', '<0xAD>', 'া', '<0xE0>', '<0xA7>', '<0xB0>', 'ত', '_', 'ম', 'ো', '<0xE0>', '<0xA7>', '<0xB0>', '_', 'ম', 'া', 'ত', '<0xE0>', '<0xA7>', '<0x83>', '<0xE0>', '<0xA6>', '<0xAD>', '<0xE0>', '<0xA7>', '<0x82>', 'ম', 'ি', '_', '<0xE0>', '<0xA6>', '<0xAD>', 'া', 'র', 'ত', '_', '<0xE0>', '<0xA6>', '<0x86>', 'ম', 'া', 'র', '_', 'ম', 'া', 'ত', '<0xE0>', '<0xA7>', '<0x83>', '<0xE0>', '<0xA6>', '<0xAD>', '<0xE0>', '<0xA7>', '<0x82>', 'ম', 'ি', '_', 'स', 'ा', 'ड', 'ा', '_', 'द', 'े', 'श', '_', 'भ', 'ा', 'र', 'त', '_', 'ह', 'ै', '_', 'भ', 'ा', 'र', 'त', '_', 'ह', 'ो', '_', 'म', '्', 'ह', 'ज', 'ो', '_', 'द', 'े', 'श', '_', 'भ', 'ा', 'र', 'त', '_', 'ह', 'म', 'र', '_', 'द', 'े', 'श', '_', 'अ', '<0xE0>', '<0xA4>', '<0x9B>', 'ि', '_', '<0xE0>', '<0xAC>', '<0xAD>', '<0xE0>', '<0xAC>', '<0xBE>', '<0xE0>', '<0xAC>', '<0xB0>', '<0xE0>', '<0xAC>', '<0xA4>', '_', '<0xE0>', '<0xAC>', '<0xB9>', '<0xE0>', '<0xAD>', '<0x87>', '<0xE0>', '<0xAC>', '<0x89>', '<0xE0>', '<0xAC>', '<0x9B>', '<0xE0>', '<0xAC>', '<0xBF>', '_', '<0xE0>', '<0xAC>', '<0xAE>', '<0xE0>', '<0xAD>', '<0x8B>', '<0xE0>', '<0xAC>', '<0xB0>', '_', '<0xE0>', '<0xAC>', '<0xAE>', '<0xE0>', '<0xAC>', '<0xBE>', '<0xE0>', '<0xAC>', '<0xA4>', '<0xE0>', '<0xAD>', '<0x83>', '<0xE0>', '<0xAC>', '<0xAD>', '<0xE0>', '<0xAD>', '<0x82>', '<0xE0>', '<0xAC>', '<0xAE>', '<0xE0>', '<0xAC>', '<0xBF>', '_', 'अ', 'स', '्', 'म', 'ा', 'क', 'ं', '_', 'द', 'े', 'श', '<0xE0>', '<0xA4>', '<0x83>', '_', 'भ', 'ा', 'र', 'त', 'व', 'र', '्', 'ष', '<0xE0>', '<0xA4>', '<0x83>']

Tokenized by mBharat tokenizer:['_ভাৰত', '_মোৰ', '_মাতৃ', 'ভূ', 'মি', '_ভারত', '_আমার', '_মাতৃ', 'ভূ', 'মি', '_सा', 'डा', '_देश', '_भारत', '_है', '_भारत', '_हो', '_म्ह', 'जो', '_देश', '_भारत', '_हम', 'र', '_देश', '_अछि', '_ଭାରତ', '_ହେଉଛି', '_ମୋର', '_ମ', 'ାତ', 'ୃ', 'ଭୂମି', '_अ', 'स्मा', 'कं', '_देश', 'ः', '_भारत', 'वर्ष', 'ः']

Figure 5: mBharat Tokenizer v/s LLaMa-2 tokenizer for input in 7 Indian languages.

in the multilingual pretraining. We grouped Bodo, Dogri, Konkani, Maithili, and Sanskrit language corpora in the same group as they follow Devanagari script, Assamese and Bangla (Bengali) were grouped together as they follow Assamese-Bengali script and Odia was kept as independent as it is typologically different than the Devanagari and Assamese-Bengali script during tokenization and pretraining. The multilingual model mParamanu was trained on the clean preprocessed pretraining data of Assamese, Bangla (Vacaspati), Bodo-Dogri (merged), Konkani, Maithili, Odia, and Sanskrit corpora as shown in Table 19.

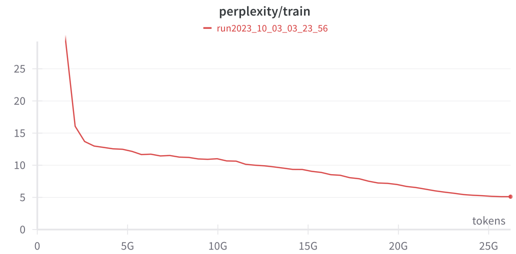
We used the same training procedure as mentioned in 3.5 but with a batch size of 32, gradient accumulation steps of 8, maximum sequence length set to 1024, i.e., 524,288 tokens per iteration and a language specific multilingual tokenizer (mBharat).

A.3.10 162M MULTILINGUAL **MPARAMANU** MODEL

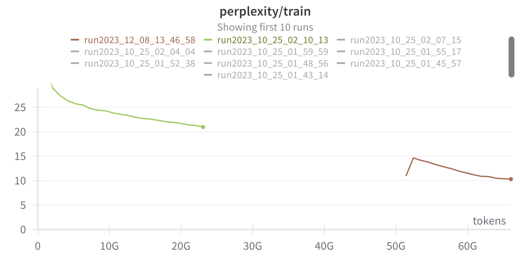
We used the same training procedure as mentioned in 3.5 but with a batch size of 32, gradient accumulation steps of 8, the maximum sequence length set to 1024, i.e., 262,144 tokens per iteration. We used language specific tokenization for the multilingual mBharat-162M model. It took 4 days and 11 hours to get trained on single A100-40G GPU with the mentioned specifications.

A.4 DATA CLEANING METHODOLOGY

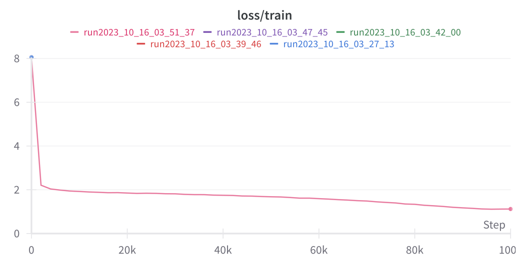
1. For languages in Devanagari, Assamese-Bengali, and Odia script (Assamese, Bangla, Hindi, Konkani, Maithili, Odia, Sanskrit), we perform splitting of sentences by punctuation "dari"/"danda" (|)
2. Removal of non literal characters for a given language and Unicode normalization
3. White space normalization.



(a) Training Perplexity for Paramanu-Bangla v/s Tokens in Billion (G=billion)



(b) Training Perplexity for Paramanu-Hindi v/s Tokens in Billion (G=billion)



(c) Training Perplexity of Paramanu-Odia v/s training steps.

(d) Training Perplexity of Paramanu-Sanskrit v/s training steps.

Figure 6: Training Perplexity v/s Tokens and Training Perplexity v/s training steps for Paramanu pretrained models.

n_params	d_model	n_layers	n_heads	dim_head	max_seq_length
15M	288	6	6	48	512
27M	512	4	8	64	1024
42M	512	8	8	64	1024
110M	768	12	12	64	1024
140M	768	15	12	64	1024
350M	1024	12	16	64	1024
425M	1280	18	10	128	1024

Table 17: Model architectures considering tokenizer size of 32000

4. Removal of English literals, Roman digits, French, German, Italian, Russian, Chinese literals and punctuation following Unicode representation using regular expressions
5. Removal of emoticons, symbols, pictographs, transport & map symbols, and iOS flags following Unicode representation using regular expression

Language	Family	Script	#Speakers
Assamese	Indo-European	Assamese-Bengali	24 M
Bangla	Indo-European	Bengali	300 M
Hindi	Indo-European	Devanagari	692 M
Konkani	Indo-European	Devanagari	2 M
Maithili	Indo-European	Devanagari	14 M
Marathi	Indo-European	Devanagari	99 M
Odia	Indo-European	Odia	43 M
Sanskrit	Indo-European	Devanagari	0.025 M
Tamil	Indo-Dravidian	Tamil	77 M
Telugu	Indo-Dravidian	Telugu	95 M

Table 18: Speaker estimates according to the Indian Census 2011

6. Removal of links, emails, HTML/XML tags, emojis, language specific punctuation, personal information like phone number, address, ID number using regular expression. We also deduplicated web scrapped pretraining corpora in respective languages

Language	**Corpus Source**	**Corpus Size**	**#Sentences**
Assamese	Indic Corp v2 + Wikipedia + Curated books (ours)	3.2 GB	5,734,166
Bangla	Vacasapati + Wikipedia + Curated books (ours)	3.6 GB	22,533,608
Hindi	IITB monolingual + Wikipedia + Curated books (ours)	15.8 GB	52,124,643
Konkani	Indic Corp v2	516.5 MB	1,337,693
Maithili	Indic Corp v2	191.3 MB	947,386
Marathi	Indic Corp v2 + Wikipedia + Curated books (ours)	12.5 GB	34,567,839
Odia	Indic Corp v2 + Wikipedia	6.2 GB	14,657,392
Sanskrit	Indic Corp v2	6.7 GB	17,034,631
Tamil	Indic Corp v2 + Wikipedia + Curated books (ours)	10.7GB	27,872,768
Telugu	Indic Corp v2 + Wikipedia + Curated books (ours)	13.5 GB	40,241,847

Table 19: Pretraining Data details after data cleaning.

A.5 Human Evaluation

A.5.1 mParamanu-162M vs Bloom 1.1B Evaluation

From Table 20 we see that Bloom 1.1B model could not distinguish languages of the same script so when we prompt Bloom with Konkani, Maithili, and Sanskrit (Devanagari script), Bloom only generated incoherent, grammatically incorrect text in Hindi whereas our multilingual model, mParamanu-162M was able to recognise prompt in respective distinguished languages Konkani, Maithili, and Sanskrit (Devanagari) to generate grammatically sound and coherent text in respective languages. Similarly, Bloom 1.1B could not identify Assamese (Assamese-Bengali Script) and generated out of context Bangla words including series of Bangla alphabets for the Assamese prompt as shown in the table 20 whereas mParamanu-162M model generated coherent sentences in Assamese when prompted with Assamese prompt. However, mParamanu-162M model is 7 times smaller in size than Bloom 1.1B and has better text generation capabilities in terms of grammar, coherence metrics for languages under consideration. This makes us to ascertain that despite Bloom being pretrained on 45 languages including Indian languages yet its text generation capabilities is very limited in terms of grammatically correct and coherent text generation for Indian languages. Therefore, its not the right way to pretrain on many languages together varying script at the same time without considering linguistics typology, features, and disproportionate pretraining low quality corpus in multilingual setting.

A.5.2 Bangla Evaluation

We hard prompted the pretrained models with the following Bangla prompts, prompt 1: ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! (phēludā tōpasēkē balala, yā lālamōhanabābukē khabara dē) – Feluda told Tapshe, Go and call Lalmohanbabu (from a popular detective novel series), prompt 2: এমন সময় হঠাৎ বাঘের ডাক! (ēmana samaẏa haṭhāt bāghēra ḍāka!) – suddenly there was a tiger roar (astonishing fearful expression), prompt 3: অপু এসে ডাকল, মা! (apu ēsē ḍākala, mā!) – Apu came and called, Mother (Apu is a character in a popular novel, also adapted into a popular movie), prompt 4: সে খুব খুশী হলো। (sē khuba khuśī halō) – she became very happy (generic statement). All these prompts are related either to common expression in Bangla or covers diverse Bangla literature and cultural context in Bangla speaking community in West Bengal, India.

Table 24 displays the human evaluation of Paramanu-Bangla 87.25M model for the mentioned Bangla prompts. We see Paramanu-Bangla 87.25M model scored an average score of 3.5/5 on grammar, 3.325/5 on coherence, 3.225/5 on creativity, and 3.2/5 on factuality metrics across top 3 generations for each Bangla prompt. Table 9 compares the performance of Paramanu-Bangla 108.5M and other LLMs including multilingual Bloom which was pretrained 45 languages including Indian languages. In this table, we can see that none of the open source LLMs have the ability

Prompt	mParamanu-162M	Bloom 1.1B
মোৰ দেশ ভাৰত (Assamese) (ōra dēśa bhārata)	মোৰ দেশ ভাৰত এখন সাংস্কৃতিক কেন্দ্ৰ (mōra dēśa bhārata ēkhana sāṁskr̥tika kēndra) মোৰ দেশ ভাৰত বিশ্বৰ বিভিন্ন ঠাইৰ সংস্কৃতিকলৈয়ে গঢ় লৈ উঠিছে (mōra dēśa bhārata biśbara bibhinna ṭhāira saṁskr̥tikalaiẏē gaṛha lai uṭhichē)	মোৰ দেশ ভাৰত (সোসাইটি বাংলাদেশ)।তথ্যসূত্র বহিঃসংযোগ বিষয়শ্রেণী:বাংলাদেশের দর্শনীয় স্থান বিষয়শ্রেণী:বাংলাদেশের প্রাচীন স্থাপত্য বিষয়শ্রেণী:বাংলাদেশের স্থাপত্য ঐতিহ্য (mōra dēśa bhārata (sōsāiṭi bāṁlādēśa)'.tathyasūtra bahiḥsaṁyōga biṣaẏaśrēṇī:bāṁlādēśēra darśanīẏa sthāna biṣaẏaśrēṇī:bāṁlādēśēra prācīna sthāpatya biṣaẏaśrēṇī:bāṁlādēśēra sthāpatya aitihya) মোৰ দেশ ভাৰতত (ক) খ) গ) ঘ) চ) ম) ক) খ) গ) ঘ) চ) ম) ক) খ) গ) ঘ) চ) ম) ক) খ) গ) ঘ) চ) ম) ক) খ) গ) ঘ) চ) ম) ক) খ) গ) ঘ) □) চ) ম) ক) খ) গ) ঘ) □) চ (mōra dēśa bhāratata (ka) kha) ga) gha) �) ca) ma) ka) kha) ga) gha) �) ca) ma) ka) kha) ga) gha) �) ca) ma) ka) kha) ga) gha) �) ca) ma) ka) kha) ga) gha) �) ca) ma) ka) kha) ga) gha) �) ca) ma) ka) kha) ga) gha) �) ca)
আমার দেশ ভারত (Bengali) āmāra dēśa bhārata	আমার দেশ ভারতবর্ষ, আমাদের দেশআমার দেশ ভারতবর্ষ হিন্দুস্থান (āmāra dēśa bhāratabarṣa, āmādēra dēśaāmāra dēśa bhāratabarṣa hindusthāna) আমার দেশ ভারতবর্ষের বিরাট অংশ এক সময় বঙ্গদেশের মত ছিল (āmāra dēśa bhāratabarṣēra birāṭa aṁśa ēka samaẏa baṅgadēśēra mata chila)	আমার দেশ ভারত, বাংলাদেশ, পাকিস্তান, শ্রীলঙ্কা, মায়ানমার, থাইল্যান্ড, ইন্দোনেশিয়া, মায়ানমার, শ্রীলঙ্কা, দক্ষিণ কোরিয়া, জাপান, চীন, জার্মানি, থাইল্যান্ড, চীন, ভারত, ইন্দোনেশিয়া, মায়ানমার, শ্রীলঙ্কা, মায়ানমার, থাইল্যান্ড, ইন্দোনেশিয়া, মায়ানমার, শ্রীলঙ্কা, দক্ষিণ কোরিয়া, জাপান, চীন, জার্মানি, থাইল্যান্ড, চীন, ভারত, ইন্দোনেশিয়া, মায়ানমার, শ্রীলঙ্কা, মায়ানমার, থাইল্যান্ড (āmāra dēśa bhārata, bāṁlādēśa, pākistāna, śrīlaṅkā, māẏānamāra, thāilyānḍa, indōnēśiẏā, māẏānamāra, śrīlaṅkā, dakṣiṇa kōriẏā, jāpāna, cīna, jārmāni, thāilyānḍa, cīna, bhārata, indōnēśiẏā, māẏānamāra, śrīlaṅkā, māẏānamāra, thāilyānḍa, indōnēśiẏā, māẏānamāra, śrīlaṅkā, dakṣiṇa kōriẏā, jāpāna, cīna, jārmāni, thāilyānḍa, cīna, bhārata, indōnēśiẏā, māẏānamāra, śrīlaṅkā, māẏānamāra, thāilyānḍa) আমার দেশ ভারত ও মুসলিম জাতির ঐক্য ও কল্যাণ কামনার সঙ্গে একাত্মতা পোষণ করার মাধ্যমে এ দেশ ও মুসলিম জাতির কল্যাণ কামনা করছি। (āmāra dēśa bhārata ō musalima jātira aikya ō kalyāṇa kāmanāra saṅgē ēkātmatā pōṣaṇa karāra mādhyamē ē dēśa ō musalima jātira kalyāṇa kāmanā karachi.)
भारत हमर देश अछि (Maithili) (bhārata hamara dēśa achi)	भारत हो म्हजो देश हो मोलांत (bhārata hō mhajō dēśa hō mōlāṁta) भारत हमर देश अछि और ओ देशक सर्वोच्च पदसँ ओतप्रोत अछि (bhārata hamara dēśa achi aura ō dēśaka sarvōcca padasam̐ ōtaprōta achi)	भारत हमर देश अछि, काहे अछि ?''हे राम !''हे राम !''हे राम !''हे राम !''हे राम ! (bhārata hamara dēśa achi, kāhē achi ?``hē rāma !''hē rāma !``hē rāma !``hē rāma !``hē rāma!) (generating Hindi) भारत हमर देश अछि। मिथुन राशि के जातक के लिए आज का दिन काफी अच्छा है। आज की दिन में नौकरी के लिए कोई परेशानी नहीं रहेगी, साथ ही आज के दिन आपके लिए कुछ अच्छा काम मिल सकता है। (bhārata hamara dēśa achi. mithuna rāśi kē jātaka kē liē āja kā dina kāphī acchā hai. āja kī dina mēṁ naukarī kē liē kōī parēśānī nahīṁ rahēgī, sātha hī āja kē dina āpakē liē kucha acchā kāma mila sakatā hai.) (generating Hindi)
ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି (Odia)	ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି ଜାତିର ଏକତ୍ରୀକରଣ ହୋଇଛି (bhārata hēuchi mōra mātr̥bhūmi jātira ēkatrīkaraṇa hōichi) ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି ଓ ମୋ ନିଜସ୍ୱତା (bhārata hēuchi mōra mātr̥bhūmi ō mō nijasvatā)	ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି। (bhārata hēuchi mōra mātr̥bhūmi.) ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି ଓ ଦେଶର ଛଟମଟ ଶୌଚାଳୟ। (bhārata hēuchi mōra mātr̥bhūmi ō dēśara chaṭamaṭa śaucāḷaẏa)
अस्माकं देशः भारतवर्षः (Sanskrit) (asmākaṁ dēśaḥ bhāratavarṣaḥ)	अस्माकं देशः भारतवर्षः सर्वातिशायी (asmākaṁ dēśaḥ bhāratavarṣaḥ sarvātiśāyī) अस्माकं देशः भारतवर्षः वर्तते (asmākaṁ dēśaḥ bhāratavarṣaḥ vartatē)	अस्माकं देशः भारतवर्षः ॥ ॥॥ (asmākaṁ dēśaḥ bhāratavarṣaḥ)

Table 20: mParamanu-162M v/s Bloom 1.1B open end text generation for Indic prompts.

Input Prompt	Text Generation
কামাখ্যা মন্দিৰ (Kamakhya Temple)	কামাখ্যা মন্দিৰত উপস্থিত হ'ব গৃহমন্ত্ৰী অমিত শ্বাহ। (kāmākhyā mandirata upasthita ha'ba gr̥hamantrī amita śbāha.) কামাখ্যা মন্দিৰ দৰ্শনৰ পাছতে মুখ্যমন্ত্ৰী হিমন্ত বিশ্ব শৰ্মাই ফ্লেগ শ্ব' মুকলি কৰিব। (kāmākhyā mandira darśanara pāchatē mukhyamantrī himanta biśba śarmāi phlēga śba' mukali kariba.) কামাখ্যা মন্দিৰলৈ যোৱা পথত যোৱা কোনো এজন লোকক ক'ভিড টেষ্ট কৰিবলৈ অনুৰোধ আৰু আহ্বান। (kāmākhyā mandiralai yōvā pathata yōvā kōnō ējana lōkakē ka'bhiḍa ṭēṣṭa karibalai anurōdha āru āhbāna.)
অসম চাহ (Assam Tea)	অসম চাহ মজদুৰ সংঘৰ মৰিগাঁও শাখাৰ ধৰ্মঘট। (asama cāha majadura saṁghara marigām̐ō śākhāra dharmaghaṭa.) অসম চাহ মজদুৰ সংঘৰ নিৰ্দেশ মৰ্মে শ্ৰমিকৰ মজুৰি ৩৫০টকালৈ বৃদ্ধি কৰাৰ দাবীত ১ নৱেম্বৰ পৰা ৩ নৱেম্বৰলৈ তিনিদিনীয়া কাৰ্যসূচীৰে বাগিছা সমূহৰ কাৰ্যবাহী পৰিচালক মজলিলাত চাহ শ্ৰমিকৰ বিভিন্ন দাবীত চতুৰ্থ বৰ্গৰ কৰ্মচাৰীৰ এক বিশাল গণ ধৰ্মঘট কাৰ্যসূচী ৰূপায়ণ কৰা হয়।। (asama cāha majadura saṁghara nirdēśa marmē śramikara majuri 350ṭakālai br̥ddhi karāra dābīta 1 navēmbara parā 3 navēmbaralai tinidinīẏā kāryasūcīrē bāgichā samūhara kāryabāhī paricālaka majalilāta cāha śramikara bibhinna dābīta caturtha bargara karmacārīra ēka biśāla gaṇa dharmaghaṭa kāryasūcī rūpāẏaṇa karā haẏa.) অসম চাহ মজদুৰ সংঘৰ যোৰহাট শাখাৰ সভাপতি ৰাজেন গোঁহাই আৰু সম্পাদক সৌৰভ গায়নে কয় যে পৰৱৰ্তী সময়ত উপযুক্ত ন্যায় ব্যৱস্থাৰ ক্ষেত্ৰত অসম চৰকাৰৰ মিনিষ্ট্ৰী অব বেংক লট অৱ অসম আৰু অসম আৰাম গৃহৰ ব্যৱস্থাৰ বাবে বহু লোকক অসম প্ৰেমী দৰিদ্ৰ ৰজাঘৰীয়া আৰ্থিক সাহাৰ্য্য প্ৰদান কৰি আহিছে। (asama cāha majadura saṁghara yōrahāṭa śākhāra sabhāpati rājēna gōm̐hāi āru sampādaka saurabha gāẏanē kaẏa yē paravartī samaẏata upayukta nyāẏa byavasthāra kṣētrata asama carakārara miniṣṭrī aba bēṁka laṭa ava asama āru asama ārāma gr̥hara byavasthāra bābē bahu lōkaka asama prēmī daridra rajāgharīẏā ārthika sāhāryya pradāna kari āhichē.)
ভূপেন হাজৰিকা (Bhupen Hazarika)	ভূপেন হাজৰিকা ন্যাস গুৱাহাটীয়ে এই ন্যাসৰ বাবে প্ৰদান কৰিছে জনজাতীয় পুৰস্কাৰ। (bhūpēna hājarikā nyāsa guvāhāṭīẏē ēi nyāsara bābē pradāna karichē janajātīẏa puraskāra.) ভূপেন হাজৰিকা হেমচন্দ্ৰ বৰুৱা অসমীয়া বিনোদন শাস্ত্ৰৰ পিতৃ স্বৰ্গীয় পিতাপত্ৰৰ স্বত্বাধিকাৰী হেমচন্দ্ৰ গোস্বামী আছিল অসমীয়া ভাষা সাহিত্যৰ এগৰাকী বিশিষ্ঠ প্ৰতিভাশালী ব্যক্তি। (bhūpēna hājarikā hēmacandra baruvā asamīẏā binōdana śāstrara pitr̥ sbargīẏa pitāpatrara sbatbādhikārī hēmacandra gōsbāmī āchila asamīẏā bhāṣā sāhityara ēgarākī biśiṣṭha pratibhāśālī byakti.) ভূপেন হাজৰিকা সেতুঘাট ৰ ওপৰত উজনি অসমত চলোৱা এক বহৎ অভিযানত বৃহৎ পৰিমাণৰ ড্ৰাগছ জব্দ কৰা হৈছে। (bhūpēna hājarikā sētughāṭa ra ōparata ujani asamata calōvā ēka bahat abhiyānata br̥hat parimāṇara ḍrāgacha jabda karā haichē.)
বিহু (Bihu)	বিহু অসমৰ এবিধ লোকনাট্যানুষ্ঠানৰ অন্যতম এক লোকনাট্যানুষ্ঠান। (bihu asamara ēbidha lōkanāṭyānuṣṭhānara anyatama ēka lōkanāṭyānuṣṭhāna.) বিহু উদযাপনৰ জৰিয়তেও মাক দেউতাকৰ উপস্থিতিত শিশুৰ বেলিকাও স্বচছলতা লাভ কৰিব পৰা যায়। (bihu udayāpanara jariẏatēō māka dēutākara upasthitita śiśura bēlikāō sbacachalatā lābha kariba parā yāẏa.) বিহু চেৰে মাজুলীৰ আকাশ বতাহ মুখৰিত। (bihu cērē mājulīra ākāśa batāha mukharita.)

Table 21: Paramanu-Assamese open end text generation for Assamese prompts capturing local context and culture.

Input Prompt	Text Generation
ଜଗନ୍ନାଥ ମନ୍ଦିର (Jagannath Mandir)	ଜଗନ୍ନାଥ ମନ୍ଦିରକୁ ଦର୍ଶନ ପାଇଁ ଆସୁଥିବା ପିଲାମାନେ ମହାପ୍ରଭୁଙ୍କ ଆଦ୍ୟ ସେବକ ଶାସନ ଗାଦିରେ ବସିଛନ୍ତି। (jagannātha mandiraku darśana pāim̐ āsuthibā pilāmānē mahāprabhuṅka ādẏa sēbaka śāsana gādirē basichanti.) ଜଗନ୍ନାଥ ମନ୍ଦିରକୁ ସଂଯୋଗ କରୁଥିବା ଜଗନ୍ନାଥ ମନ୍ଦିରର ତିହ୍ୟ ସ୍ଥାନର ବିକାଶ ପାଇଁ ୫କୋଟି ଟଙ୍କା ଅନୁଦାନ ଦିଆଯାଇଥିବାବେଳେ ବସ୍ ଷ୍ଟାଣ୍ଡର ନବୀକରଣ ନିର୍ମାଣ କରାଯାଇ। (jagannātha mandiraku saṁyōga karuthibā jagannātha mandirara tihẏa sthānara bikāśa pāim̐ 5kōṭi ṭaṅkā anudāna diāyāithibābēḷē bas ṣṭāṇḍara nabīkaraṇa nirmāṇa karāyāi.) ଜଗନ୍ନାଥ ମନ୍ଦିର ପରିଚାଳନା କମିଟିର ପୂର୍ବତନ ସଦସ୍ୟ ହେଉଛନ୍ତି ଅନନ୍ତ କୁମାର ଜେନା। (jagannātha mandira paricāḷanā kamiṭira pūrbatana sadasẏa hēuchanti ananta kumāra jēnā.)
ପୁରୀ ସାଗର ବେଳାଭୂମି (Puri Sea Beach)	ପୁରୀ ସାଗର ବେଳାଭୂମିରେ ଶୁଭାରମ୍ଭ ହୋଇଛି ଆନ୍ତର୍ଜାତୀୟ ଖ୍ୟାତି ସମ୍ପର୍ଣ୍ଣ ବାଲୁକା ଶିଳ୍ପୀ ସୁଦର୍ଶନ ପଟ୍ଟନାୟକଙ୍କ ଦ୍ୱାରା ବିଶ୍ୱ ବାଲୁକା କୃତୀ। (purī sāgara bēḷābhūmirē śubhārambha hōichi āntarjātīẏa khẏāti samparṇṇa bālukā śiḷpī sudarśana paṭṭanāẏakaṅka dvārā biśva bālukā kr̥t.) ପୁରୀ ସାଗର ବେଳାଭୂମି ମହୋତ୍ସବର ପ୍ରଥମ ଦିନରେ ପୁରୀ ଶ୍ରୀଜଗନ୍ନାଥ ମନ୍ଦିରରେ ଆକର୍ଷଣୀୟ ନୃତ୍ୟକଳା ଦୃଶ୍ୟ। (purī sāgara bēḷābhūmi mahōtsabara prathama dinarē purī śrījagannātha mandirarē ākarṣaṇīẏa nr̥tẏakaḷā dr̥śẏa.) ପୁରୀ ସାଗର ବେଳାଭୂମିରେ ୩୨ ଫୁଟର ପାନ୍ଥେରା। (purī sāgara bēḷābhūmirē 32 phuṭara pānthērā.)
ଚିଲିକା ହ୍ରଦ (Chilka Lake)	ଚିଲିକା ହ୍ରଦ ଭିତରକନିକା ପ୍ରଭୃତିରେ ବିଭିନ୍ନ ପ୍ରଜାତିର ଡଲଫିନ ମିଳୁଥିଲା। (cilikā hrada bhitarakanikā prabhr̥tirē bibhinna prajātira ḍalaphina miḷuthilā.) ଚିଲିକା ହ୍ରଦ ଭୂଇଁରେ ଛୁଇଁବ ସମୁଦ୍ର ଆସିଲା ସାମାନ୍ୟ ବହଳିଆ ଜୁଆର। (cilikā hrada bhūim̐rē chuim̐ba samudra āsilā sāmānẏa bahaḷiā juāra.) ଚିଲିକା ହ୍ରଦ ବ୍ରାହ୍ମଣୀ ବୈତରଣୀ ଓ ଅନ୍ୟ ଜଳଦ୍ୱୀପ ଦେଇ ଏହି ନଦୀ ପୁରୀକୁ ପ୍ରଜନନ କରେ ଓ ବ୍ରାହ୍ମଣୀ ନଦୀର ଜଳ ଯାଜପୁର ଟିକାୟତନନ୍ଦପୁରରେ ପ୍ରସ୍ତାବିତ ରାସପଥର ପୁନରୁଦ୍ଧାର କରି ଏ ଜିଲ୍ଲା ପ୍ରଗତିର ସୂତ୍ରପାତ ହୁଏ। (cilikā hrada brāhmaṇī baitaraṇī ō anẏa jaḷadvīpa dēi ēhi nadī purīku prajanana karē ō brāhmaṇī nadīra jaḷa yājapura ṭikāẏatanandapurarē prastābita rāsapathara punaruddhāra kari ē jillā pragatira sutrapāta huē.)
କୋଣାର୍କ ସୂର୍ଯ୍ୟ ମନ୍ଦିର (Konark Sun Temple)	କୋଣାର୍କ ସୂର୍ଯ୍ୟ ମନ୍ଦିରକୁ ବିଶ୍ୱ ତିହ୍ୟର ମାନ୍ୟତା ଦାବିରେ ମୁଖ୍ୟମନ୍ତ୍ରୀଙ୍କୁ ଚିଠି ଲେଖିଲେ କୋଣାର୍କ ବିଧାୟକ। (kōṇārka sūryẏa mandiraku biśva tihẏara mānẏatā dābirē mukhẏamantrīṅku ciṭhi lēkhilē kōṇārka bidhāẏaka.) କୋଣାର୍କ ସୂର୍ଯ୍ୟ ମନ୍ଦିର ପାଦଦେଶରେ କୋଣାର୍କ ମହୋତ୍ସବ ଉଦ୍ ଘାଟିତ। (kōṇārka sūryẏa mandira pādadēśarē kōṇārka mahōtsaba ud ghāṭita.) କୋଣାର୍କ ସୂର୍ଯ୍ୟ ମନ୍ଦିରର ବିଭିନ୍ନ ଅବ୍ୟବସ୍ଥା ନେଇ କୋଣାର୍କ ଯାତ୍ରୀ ପ୍ରତିବାଦ କରିବା ପରେ ବିଦେଶୀ ଯାତ୍ରୀଙ୍କ ଆଗମନ ଉପରେ ପ୍ରଶାସନ ୧୪୪ ଧାରା ଲାଗୁ କରିଥିଲା। (kōṇārka sūryẏa mandirara bibhinna abẏabasthā nēi kōṇārka yātrī pratibāda karibā parē bidēśī yātrīṅka āgamana uparē praśāsana 144 dhārā lāgu karithilā.)

Table 22: Paramanu-Odia's text generation for 4 Odia prompts capturing local context and culture.

Language	Input Prompt	Text Generation
Assamese	ভাৰত মোৰ মাতৃভূমি	মোৰ দেশ ভাৰত বিশ্বৰ বিভিন্ন ঠাইৰ সংস্কৃতিকলৈয়ে গঢ় লৈ উঠিছে। মোৰ দেশ ভাৰত এখন সাংস্কৃতিক কেন্দ্ৰ।
Bangla	ভারত আমার মাতৃভূমি	আমার দেশ ভারতবর্ষ আমাদের দেশ । আমার দেশ ভারতবর্ষের বিরাট অংশ এক সময় বঙ্গদেশের মত ছিল।
Dogri	साडा देश भारत है	साडा देश भारत है हाँ आज के समय से आने वाला उत्तर भारतीय है जहां धीरूभाई अम्बानी फॉर युवा का मुजफ्फरपुर का लोग है वहां के बावजूद बिग बॉस नाम की तरफ से भी कुछ लकी।
Konkani	भारत हो म्हजो देश	भारत हो म्हजो देश ह्या राष्ट्राची वाट लायल्या भारत हो म्हजो देश भारत म्हज्या इतिहासांत देव म्हज्या इतिहासांत आस्पावता आदलें पैकूच आसा त्या इतिहासाचो आदलो काळ आशिल्लो हाका इतिहासाचो चवथो भाग म्हण्टा।
Maithili	भारत हमर देश अछि	भारत हमर देश अछि और ओ देशक सर्वोच्च पदसँ ओतप्रोत अछि भारत हो म्हजो देश हो मोलांत।
Odia	ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି	ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି ଜାତିର ଏକତ୍ରୀକରଣ ହୋଇଛି ଭାରତ ହେଉଛି ମୋର ମାତୃଭୂମି ଓ ମୋ ନିଜସ୍ୱତା
Sanskrit	अस्माकं देशः भारतवर्षः	अस्माकं देशः भारतवर्षः सर्वातिशायी। अस्माकं देशः भारतवर्षः इति ख्यातः।

Table 23: mParamanu 162 M pretrained model text generation

Prompt	Text Generation (top-3)	Grammar	Consistency	Coherency	Factuality
ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে!	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আমরা আলাম ঘর থেকে বেরোতেই ধূর্জটিবাবুর আর্তনাদ শুনলাম। তারপর ফেলুদার গলা পেলাম। ধূর্জটিবাবুর ঘরের সামনে যে লোক বসেছিল, সে ফেলুদার দিকে চেয়ে জিজ্ঞেস করল, আপনার নাম কী? ফেলুদা আর্তনাদ শুনে ফেলুদা বাইরে চলে গেল। আমি আরও মিনিটখানেক বাইরে হাঁটলাম। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! উনি আসছেন। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! ভদ্রলোক জিগ্যেস করবেন, আর সত্যি বলতে কী, আমার সন্দেহ আছে। বাইফোকালসাই।–তাও অবিশ্যি উনি ছাড়।	3/0.66	2.8/0.53	2.6/0.4	2.5/0.33
এমন সময় হঠাৎ বাঘের ডাক!	এমন সময় হঠাৎ বাঘের ডাক! বাঘের সে ডাক বাঘের স্বর নয়! দূরে গিরিগঞ্জের ধানখেতের পিছন হইতে যেন কাঁপিতে কাঁপিতে ছুটিয়া আসিতেছে। এমন সময় হঠাৎ বাঘের ডাক! সন্ত ভয়ে চী- চী করিয়া উঠিল। মনে হইল যেন সেই বনের হরিণগুলা আবার জঙ্গলের দিক হইতে আসিয়া উপস্থিত হইল। এমন সময় হঠাৎ বাঘের ডাক! আমার ভয় পাবার হয়তো কিছু ছিল। নইলে দুবে সকলের চোখের ওপরে একটা শজারুর মতো বড়ো চোখা শিং অনেকদিন আগেই মেরেছি কি-না? না, না, আমি মিথ্যে ভয় করি না। সে গলার আওয়াজ এবার কাছেই কোথায় মিলিয়ে গেল!	3.3/0.60	2.9/0.40	2.8/0.30	2.8/0.3
অপু এসে ডাকল, মা!	অপু এসে ডাকল, মা, খেতে দাও। অপু এসে ডাকল, মা, ও মা, কোথা যাও? অপু এসে ডাকল, মা, শিগগির এসো, গ্রামের ওদিক থেকে ডাকছে।	4	4	4	4
সে খুব খুশী হলো।	সে খুব খুশী হলো। সবাই তাকে এত সম্মানের চোখে দেখে। প্রথম দিন থেকে দেখে এবং অভ্যর্থনা পাওয়ার পর তার প্রতি শ্রদ্ধা জানায়। সে খুব খুশী হলো। ভালো হয়েছে, মা। সে খুব খুশী হলো। সবকিছু তার খোলসা হয়ে গেছে।	3.7/0.70	3.6/0.6	3.5/0.50	3.5/0.50

Table 24: Human evaluation results of open-end text generation of Paramanu-Bangla 87.25M for 4 Bangla prompts on Grammar, Coherency, Creativity, and Factuality metrics. We use the format (a/b) where a is the average score of top 5 generations on scale of 0 (worst) to 5 (best) and b is the normalized score according to a; $b = a_{min}/(a_{max} - a_{min})$

to generate grammatically, coherent sentences in Bangla except the Bloom series keeping aside the factuality. Our monolingual model, Paramanu-Bangla 108.5M model has scored 202.7% better and 166.6% better on grammar metric, 254.84 % and 292.86% on coherence metric, 182.22% and 222.53% on creativity metric, 159.4% and 219.23% on factuality metric than Bloom 3B, and Bloom 7B models respectively, despite Paramanu-Bangla is 28.3 times and 67 times smaller in terms of total number of parameters than Bloom 3B and Bloom 7B. Interestingly, we found that Bloom 560M model performed better on all the metrics than the bigger Blooom models. Here, we can see the curse of multilinguality coming in picture as the increase of number of parameters of the models with 45 languages in the pretraining data seems to downgrade the text generation capabilities. Among other series, GPT-Neo 1.3B tends to be better than LLaMa-2 7B, OPT 6.7B and GPT2-XL for Bangla text generation. Table 26 compares the text generation of OPT 6.7B model with our pretrained Paramanu-Bangla model for Bangla open end text generation. Table 27 compares the text generation of LLaMa-2 7B with our Paramanu-Bangla. Table 28 compares GPTJ 6B with Paramanu-Bangla, and table 30 compares the series of multilingual Bloom models with Paramanu-Bangla and table 29 compares the GPT2-XL, GPT Neo 1.3B with our Paramanu-Bangla for open end text generation for the above mentioned prompts. We observed that GPTNeo 1.3B even generated Arabic text when hard prompted with Bangla prompt এমন সময় হঠাৎ বাঘের ডাক! (ēmana samaẏa haṭhāt bāghēra ḍāka!)

From Figures 12 and 13 we can see that GPT-3.5 Turbo mixed Bengali with Assamese languages together as a response to Bengali prompt. The mixed text generation make no sense in neither Bangla nor Assamese. When we hard prompted GPT-3.5 Turbo through Open AI website, we received the same output to our prompts. Therefore, our evaluators scored top 3 responses with the same scores to 4 Bangla prompts.

A.5.3 Sanskrit Evaluation

We hard prompted the models with the following prompts: अस्माकं देशः भारतवर्षः (asmākaṁ dēśaḥ bhāratavarṣaḥ) – our country Bharatavarsha, वेदः चत्वारः सन्ति (vēdaḥ catvāraḥ santi) – there are four Vedas, मह्यं मिष्टान्नं रोचते (mahyaṁ miṣṭānnaṁ rōcatē) – I like sweets, and किमर्थं त्वं गच्छसि (kimarthaṁ tvaṁ gacchasi) – why are you going.

Table 11 compares the performance of mParamanu-162M and other LLMs including multilingual Bloom which was pretrained on Indian languages. We can see that none of the LLMs have the ability to generate grammatically, coherent sentences in Sanskrit keeping aside the factuality. Our multilingual model, mParamanu-162M has scored the highest among all on grammar (3.75/5), coherence (3.166/5), creativity (2.166/5), and factuality (1.75/5) whereas Bloom 3B scored 0.166/5 on grammar, 0.0833/5 on coherence, and 0/5 for both creativity and factuality metrics respectively. GPT-3.5-Turbo (ChatGPT) has scored very poorly 0.25/5 on grammar & coherence metrics, 0.1818/5 in creativity and 0.33/5 on factuality metrics respectively for Sanskrit text generation. mParamanu-162M is smaller by 44.25 times compared to 7B LLaMa-2 model and yet it has shown its high quality text generation in Sanskrit than ChatGPT, LLaMa, and Bloom series of models.

Table 31 compares the text generation of OPT 6.7B model with our pretrained mParamanu model for Sanskrit open end text generation. Table 32 compares the text generation of LLaMa-2 7B with our mParamanu. Table 33 compares GPTJ 6B with mParamanu, and table 35 compares the series of multilingual Bloom models with mParamanu and Table 34 compares the GPT2-XL, GPT Neo 1.3B with mParamanu for open end text generation. Figure 10 and Figure 11 are GPT-3.5 Turbo responses to respective Sanskrit prompts.

A.5.4 HINDI EVALUATION

We hard prompted the LLMs (LLaMa-2, Bloom Series, GPTNeo 1.3B, GPT2-XL) and our Paramanu-Hindi 162M pretrained model with the following Hindi prompts, prompt 1: सचिन तेंदुलकर (Sachin Tendulkar), prompt 2: शाहरुख़ ख़ान (Shah Rukh Khan), prompt 3: महात्मा गांधी (Mahatma Gandhi), and prompt 4: लता मंगेशकर (Lata Mangeshkar). These prompts are related to popular celebrities across cricket, films, politics and music respectively in India.

Table 10 compares the performance of Paramanu-Hindi 162M and other LLMs including multilingual Bloom which was pretrained on Indian languages. In this table, we can complete see that none of the open source LLMs have the ability to generate grammatically, coherent sentences in Hindi except the Bloom series. Our monolingual model, Paramanu-Hindi 162M has performed better by 17.25% on grammar, by 46.05% on coherence, by 62.5% on creativity, and by 238.5% on factuality compared to Bloom 3B model despite being 19 times smaller in size. Table 37 compares the text generation of OPT 6.7B model with our pretrained Paramanu-Hindi model for Hindi open end text generation. Table 38 compares the text generation of LLaMa-2 7B with our Paramanu-Hindi 162M. Table 39 compares GPT-J 6B with Paramanu-Hindi 162M, and Table 40 compares the series of multilingual Bloom models with Paramanu-Hindi 162M and Table 36 compares the GPT2-XL, GPT Neo 1.3B with our Paramanu-Hindi 162M for open end text generation. We observed GPT-J 6B generated random text in Japanese too when prompted with Hindi prompt शाहरुख़ ख़ान (Shah Rukh Khan) and also generated random text in Portuguese when prompted with लता मंगेशकर (Lata Mangeshkar), and text in Kannada and Serbian when prompted with सचिन तेंदुलकर (Sachin Tendulkar). None of these LLMs (LLaMa-2 7B, OPT 6.7B, GPT-J 6B, GPTNeo 1.3B, and GPT2-XL have the ability to generate text in Hindi and can not generalize beyond English or some European languages.

A.5.5 ASSAMESE AND ODIA

For both Assamese, and Odia, we were not able to perform human evaluation due to lack of resources at our end. However, we yet prompted our models with local cultural prompts. For Assamese, we prompted with কামাখ্যা মন্দিৰ (Kamakhya Temple) – a very popular temple in Assam, অসম চাহ (Assam Tea), ভূপেন হাজৰিকা (Bhupen Hazarika) – a popular singer, and বিহু (Bihu) – the biggest Assamese festival. Table 21 shows the generated output from our Paramanu-Assamese.

For Odia, we prompted with ଜଗନ୍ନାଥ ମନ୍ଦିର (Jagannath Mandir) – a renowed temple in Odisha, ପୁରୀ ସାଗର ବେଳାଭୂମି (Puri Sea Beach), ଚିଲିକା ହ୍ରଦ (Chilka Lake) – the biggest lake in India, and କୋଣାର୍କ ସୂର୍ଯ୍ୟ ମନ୍ଦିର (Konark Sun Temple) – an ancient Sun temple and UNESCO World Heritage site. Table 22 lists down the responses of our Paramanu-Odia for the given prompts.

Based on Google Translate, we found the text generation from both of our models captured local context, historical and factual responses related to the cultural and local prompts used to query the models. Table 20 shows the results. We also observed that multilingual Bloom series is unable to distinguish languages of similar script so when we prompted Bloom with Assamese prompt, Bloom only generated text in Bangla whereas our multilingual model, mParamanu has the ability to distinguish languages of the same script unlike Bloom.

B PARAMANU INSTRUCTION-TUNED MODELS

B.1 PARAMANU-BANGLA-INSTRUCT AKA (BANGLA-GPT)

We instruction tuned our Paramanu-Bangla pretrained model on 23k Bangla instructions. We named instruction-tuned Bangla model as Paramanu-Bangla-instruct aka (Paramanu-Bangla as shown in the figures)

Fig 8 shows the high quality text generation capabilities of our pretrained Bangla model for query: আমার জীবন (āmāra jībana, my life) and Fig 9 for query: দেশের রাজনৈতিক অস্থিরতা (dēśēra rājanaitika asthiratā, political instability in the country). The Bangla model has been consistent in generating high quality grammatically correct, coherent sentences.

Fig 14 and Fig 15 exhibit the in-context learning of our Bangla model that it can also do grammar correction without extra fine-tuning. Fig 16 shows that our model can also write grocery list in order to prepare delicious Bengali food. Fig 17 answers question related to archaeological history. Fig 18 answers in details about benefits of yoga practice. Fig 19 shows the poem writing skills of our Bangla model. Fig 20 answers question related to finance domain. Fig 21 shows that our model can even write in the style of great Bengali Nobel laureate poem, Rabindranath Tagore. Fig 22 that our model can also write food recipes. Fig 23 and Fig 24 show that our model can also answer questions from education domain related to benefits of education, sex education, etc. Fig 25 answers a question from the sports domain. Fig 26 and Fig 27 show the amazing capability of our 108.5M model that it can write long stories of two pages being grammatically correct, coherent, creative and consistent. Fig 28 show itemized response to a question.

All these demonstrations show the various tasks execution capabilities of our model despite our model is just 108.5M in size but its very powerful and the first Bangla generative model of such kind exclusively pretrained on Bangla corpus and instruction tuned on 23k Bangla instructions.

B.2 PARAMANU-HINDI-INSTRUCT AKA (HINDI-GPT)

Fig 29 shows that our instruction tuned Hindi model has been able to answer user's question from healthcare domain in detailed manner discussing how to get good sleep in the night without any repetition and inconsistency. Fig 30 answers the question from public administration. Fig 31 plans an travel itinerary for vacation. Fig 32 shows that our model can even write about Indian recipes for healthy food. Fig 33 shows that our model can also answer questions from finance and technology. Our model can also answers questions, as shown in Fig 34. Fig 35 shows that our Hindi model can also design a lecture course for students summarizing in brief about the content of the chapters. Fig 36 takes a reading comprehension passage and a question as input and answers the relevant answer to the question.

All these demonstrations show the various tasks execution capabilities of our model despite our model is just 367.5M in size but its very powerful and the first Hindi generative model of such kind exclusively pretrained on Hindi corpus and instruction tuned on 23k Hindi instructions.

B.3 PARAMANU-TAMIL-INSTRUCT AKA (TAMIL-GPT)

Fig 39 answers student career related question. Fig 40, Fig 41, Fig 42, Fig 43 and Fig 44 display the various instructions following capabilities such as domain knowledge in politics and civics, climate, national parks to human life related queries, etc. of Paramanu-Tamil-instruct model.

B.4 PARAMANU-TELUGU-INSTRUCT AKA (TELUGU-GPT)

Fig 45, Fig 46, Fig 47, Fig 48 and Fig 49 demonstrate various instruction following capabilities of Telugu model and knowledge in various domains from social sciences to geology to movie celebrities and more.

B.5 Inference Speed on CPU

Table 25 shows the inference speed (tokens/sec) of our models in FP32 precision with float32 forward pass and the entire calculation of the forward pass is kept in FP32. As we see that as we keep increasing the number of parameters in the model, the inference speed gets lower which is not preferable for deployment purpose due to larger model checkpoint files (4 bytes per every individual weight) and forward pass is relatively slow. One of the common inference optimization employed in practice is to quantize the model parameters to lower precision, while slightly giving up on precision (correctness) in return for smaller checkpoint sizes and faster forward passes (as most of the inference uses integer arithmetic). Table 25 shows the CPU inference speed of our models without any quantization.

Model	**Inference Speed in FP32**
Paramanu-Assamese	80.4732
Paramanu-Bangla	24.3267
Paramanu-Hindi 367.5M	12.9057
Konkani-Maithili GPT	160.8750
mParamanu 162M	12.7106
Paramanu-Marathi	24.8750
Paramanu-Odia	24.5353
Paramanu-Sanskrit	22.6757
Paramanu-Tamil	24.5353
Paramanu-Telugu	24.1245

Table 25: CPU inference speed (tokens/sec) of models in FP32 precision.

C Background

C.0.1 Language Modeling

This objective of the language modeling can be formally described as maximizing the probability of a sequence of tokens $w_1, w_2, \ldots, w_N$

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \ldots, w_{i-1})$$

where $p(w_t|w_0, \ldots w_{t-1})$ is the probability of token w_t given the sequence of previous tokens $w_0, \ldots, w_{t-1}$.

The performance of a language model is generally evaluated using the total *cross-entropy loss*, i.e, the negative log-likelihood of the observed data under the model under consideration, which for a given dataset is defined as:

$$Loss = -\frac{1}{N}\sum_{i=1}^{n} \log(P(w_i \mid w_1, w_2, \ldots, w_{i-1}))$$

Lower the loss better is the model; however, just computing the loss may not be intuitive. Therefore, *Perplexity* is a metric to evaluate the performance of a given language model which is the exponent of the average loss.

$$Perplexity = \exp{(Loss)}$$

C.0.2 Fertility Score of Tokenizers

The fertility score is a key metric used to evaluate the performance of tokenizers in natural language processing (NLP). It quantifies how well a tokenizer divides a given text into meaningful tokens while preserving the linguistic structure and semantic content of the original input. Specifically, the fertility score is defined as the ratio of the number of tokens produced by the tokenizer to the expected number of tokens that would be produced by an idealized, reference tokenization. Mathematically, it is expressed as:

$$F = \frac{N_{\text{tokens}}}{N_{\text{expected}}}$$

Where N_{tokens} is the number of tokens generated by the tokenizer, and N_{expected} is the number of tokens expected in an ideal scenario, often determined through linguistic analysis or human annotations. A fertility score close to 1 indicates that the tokenizer is performing optimally, generating a number of tokens that aligns with the reference standard. A score greater than 1 suggests over-segmentation, where the tokenizer generates more tokens than necessary, possibly losing some meaning or context in the process. A score below 1 indicates under-segmentation, where the tokenizer groups multiple linguistic units into fewer tokens, potentially missing out on finer details.

Several factors can affect the fertility score of a tokenizer, including the granularity of tokenization (e.g., character-level vs. word-level), vocabulary coverage (whether the tokenizer can match entire words or subword units), and how special characters and punctuation are handled. Moreover, the characteristics of the language being processedsuch as morphological complexitycan also influence tokenization, especially in languages with rich inflections or compounds.

The fertility score is important for ensuring that tokenization strikes a balance between semantic precision and computational efficiency. In tasks like machine translation, where retaining meaning is crucial, a higher fertility score may be desired to preserve linguistic nuances. However, in real-time applications where speed is critical, a lower fertility score may be preferred to reduce the number of tokens and computational load.

For example, consider the sentence “I love natural language processing.” A tokenizer that produces the following tokens: `[I, love, natural, language, processing]` would have a fertility score of:

$$F = \frac{5}{5} = 1$$

This score suggests that the tokenizer is performing as expected. However, if another tokenizer splits “natura” into “natur” and “al,” producing the tokens `[I, love, natur, al, language, processing]`, the fertility score would be:

$$F = \frac{6}{5} = 1.2$$

This indicates over-segmentation, with more tokens than ideal.

In summary, the fertility score of a tokenizer is an essential metric that helps to assess the efficiency and effectiveness of tokenization strategies. By optimizing the fertility score, one can ensure that tokenization maintains the right balance between computational efficiency and the preservation of meaning, making it a crucial aspect of NLP systems.

C.0.3 Rotary Position Embedding (RoPE)

Transformer-based models rely on positional embeddings to encode position and relative location information of words in a text. *Rotary Position Embedding (RoPE)* is a position encoding technique proposed by (Black et al., 2022). Instead of adding positional embeddings or relative positional embeddings to token embeddings, RoPE rotates the token embedding by a fixed factor (θ) in the higher-dimensional space to encode relative positional embeddings. In other words, RoPE encodes the absolute positions with a rotation matrix and meanwhile incorporates the explicit relative position dependency in self-attention formulation. The intuition behind RoPE is that we can represent the token embeddings as complex numbers and their positions as pure rotations that we apply to them. If we shift both the query and key by the same amount, changing absolute position but not relative position, this will lead both representations to be additionally rotated in the same manner. Thus, the angle between them will remain unchanged and, thus, the dot product will also remain unchanged. By exploiting the nature of rotations, the dot product used in self-attention will have the property for preserving relative positional information while discarding absolute position.

C.0.4 ROOT MEAN SQUARE NORMALIZATION (RMSNORM)

To improve the training stability, some LLMs (Chinchilla (Hoffmann et al., 2022a), LLaMa (Touvron et al., 2023)) have normalized the input of each transformer sub-layer, instead of normalizing the output using RMSNorm normalizing function as introduced by (**?**). *RMSNorm* normalizes the activations based on their root mean square (RMS) value instead of normalizing the inputs based on their mean and variance.

RMSNorm accelerates the training and inference with similar performance in these large models. It is reported that replacing LayerNorm (Ba et al., 2016) with RMSNorm can achieve comparable performance and improve training and inference time by 7-64%. Narang et al. (2021) showed that RMSNorm improves the pre-training speed by 5% compared with the LayerNorm baseline.

C.1 MODEL FLOPS UTILIZATION (MFU)

Model FLOPs Utilization (MFU) Chowdhery et al. (2023) estimate is the ratio of the observed throughput (tokens-per-second) relative to the theoretical maximum throughput of a system at peak FLOPs. Model flops utilization (MFU) estimate the number of flops (floating point operations) done per iteration. It quantifies how efficiently the GPUs are utilized in model training.

C.2 MAXIMAL UPDATE PARAMETERIZATION

As the size of large language models (LLMs) and the scale of the dataset used in pretraining are expensively large, it is not feasible to perform hyperparameter tuning in LLMs. Yang et al. (2021) used a technique called maximal update parameterization (μP) to transfer the hyperparameters learnt from tuning of a small model to a larger model and found that the optimal hyperparameter values become stable across neural network sizes when the models have been parameterized using (μP).

C.3 CARBON FOOTPRINT

To measure carbon footprint for our pretraining, we follow Touvron et al. (2023):

$$tCO_2eq = MWh \times 0.385$$

The power consumption can be calculated as

$$Wh = \text{GPU-hours} \times (\text{GPU power consumption}) \times \text{PUE}$$

where PUE is Power Usage Effectiveness.

We observed during pretraining that our single A100 40G consumes 250 Watt consistently.

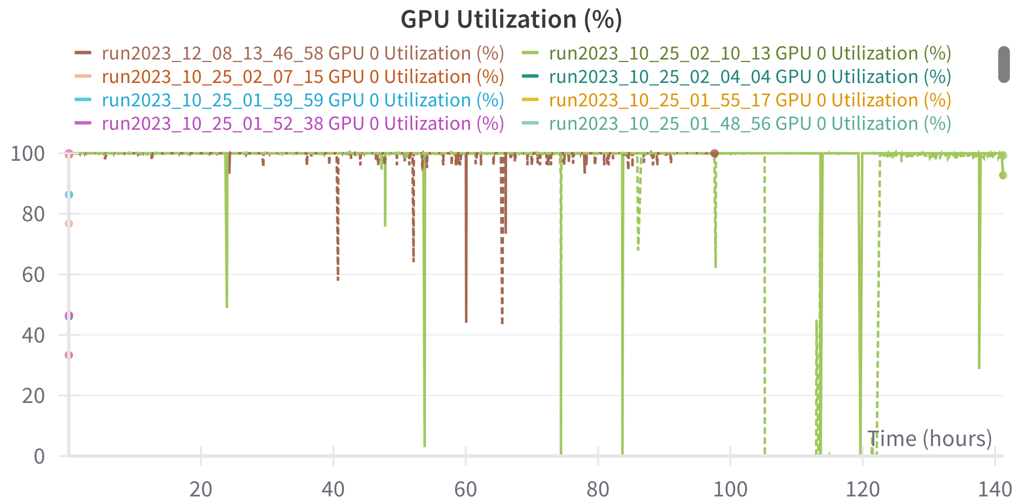


Figure 7: GPU Utilization (%) for pretraining Hindi language model from scrtach.

Figure 8: Prompting our pretrained Paramanu-Bangla 108.5M model with query আমার জীবন (āmāra jībana) (My life)

Figure 9: Prompting our pretrained Paramanu-Bangla model with query: দেশের রাজনৈতিক অস্থিরতা (dēśēra rājanaitika asthiratā.)

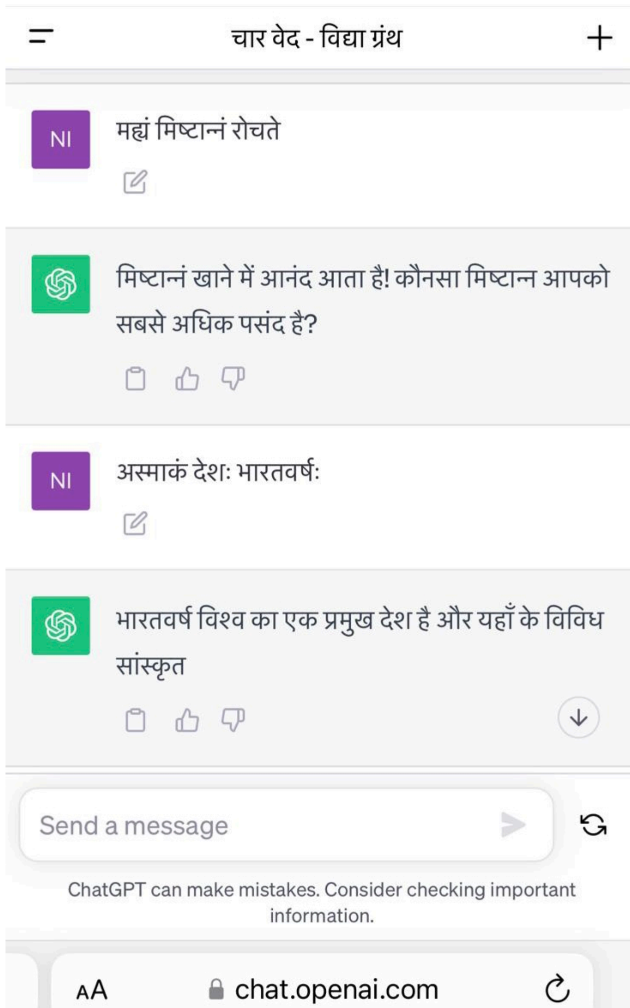


Figure 10: GPT-3.5 Turbo response to Sanskrit prompt: अस्माकं देशः भारतवर्षः (asmākaṁ dēśaḥ bhāratavarṣaḥ), मह्यं मिष्टान्नं रोचते (mahyaṁ miṣṭānnaṁ rōcatē.) Accessed in October 2023.

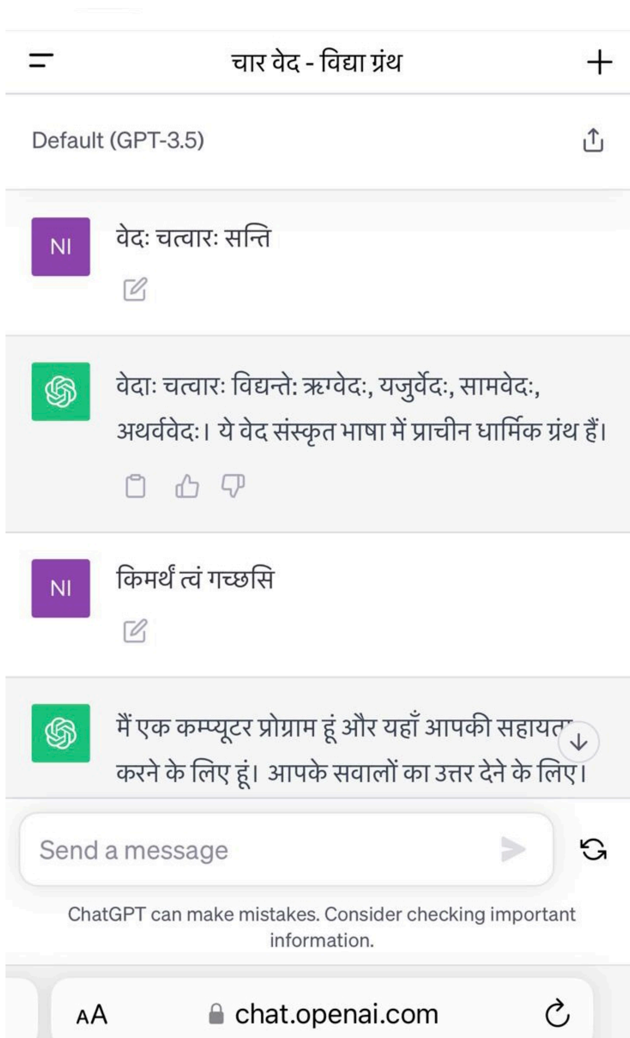


Figure 11: GPT-3.5 Turbo response to Sanskrit prompt: वेदः चत्वारः सन्ति (vēdaḥ catvāraḥ santi), किमर्थं त्वं गच्छसि (kimarthaṁ tvaṁ gacchasi.) Accessed in October 2023.

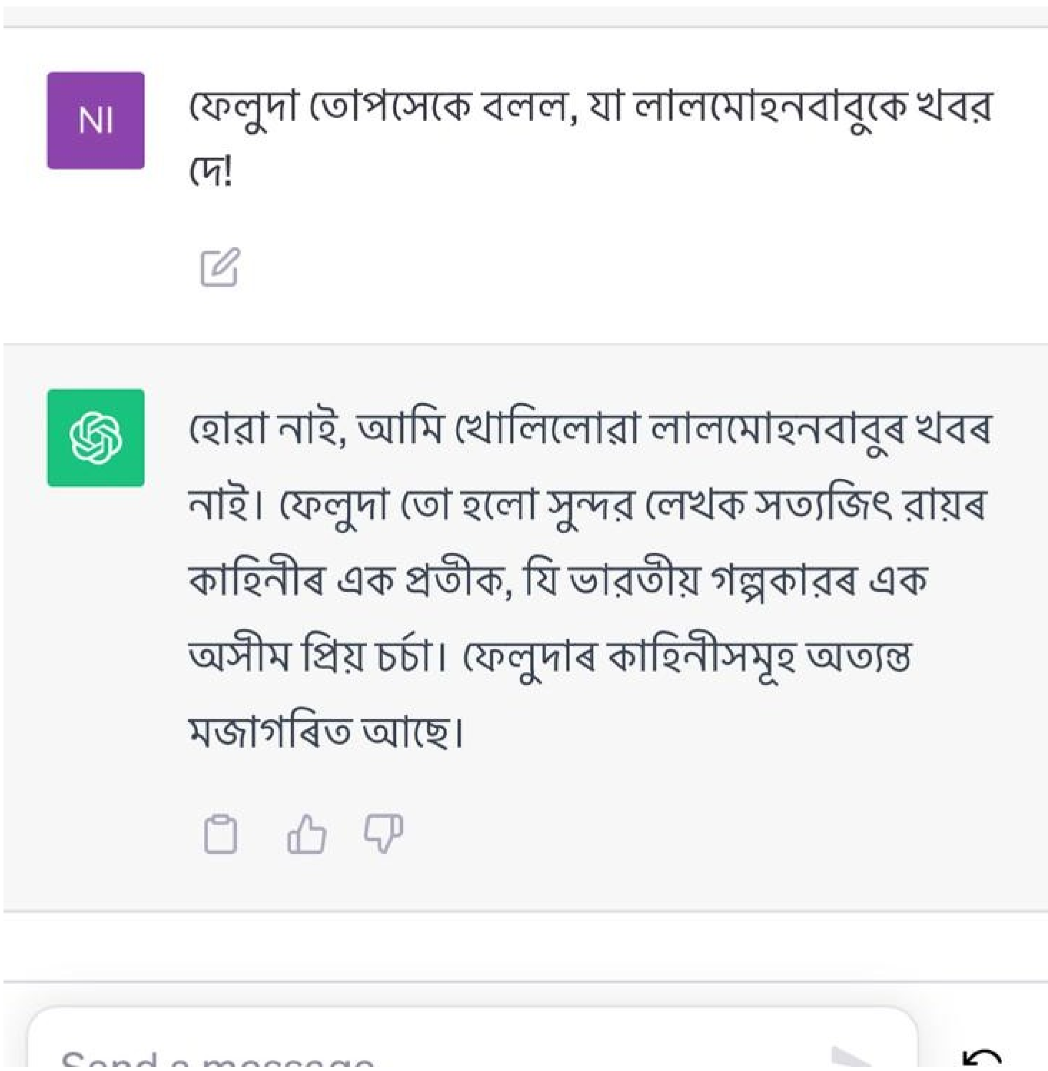


Figure 12: GPT-3.5 Turbo response to Bangla prompt ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! Accessed in October 2023.

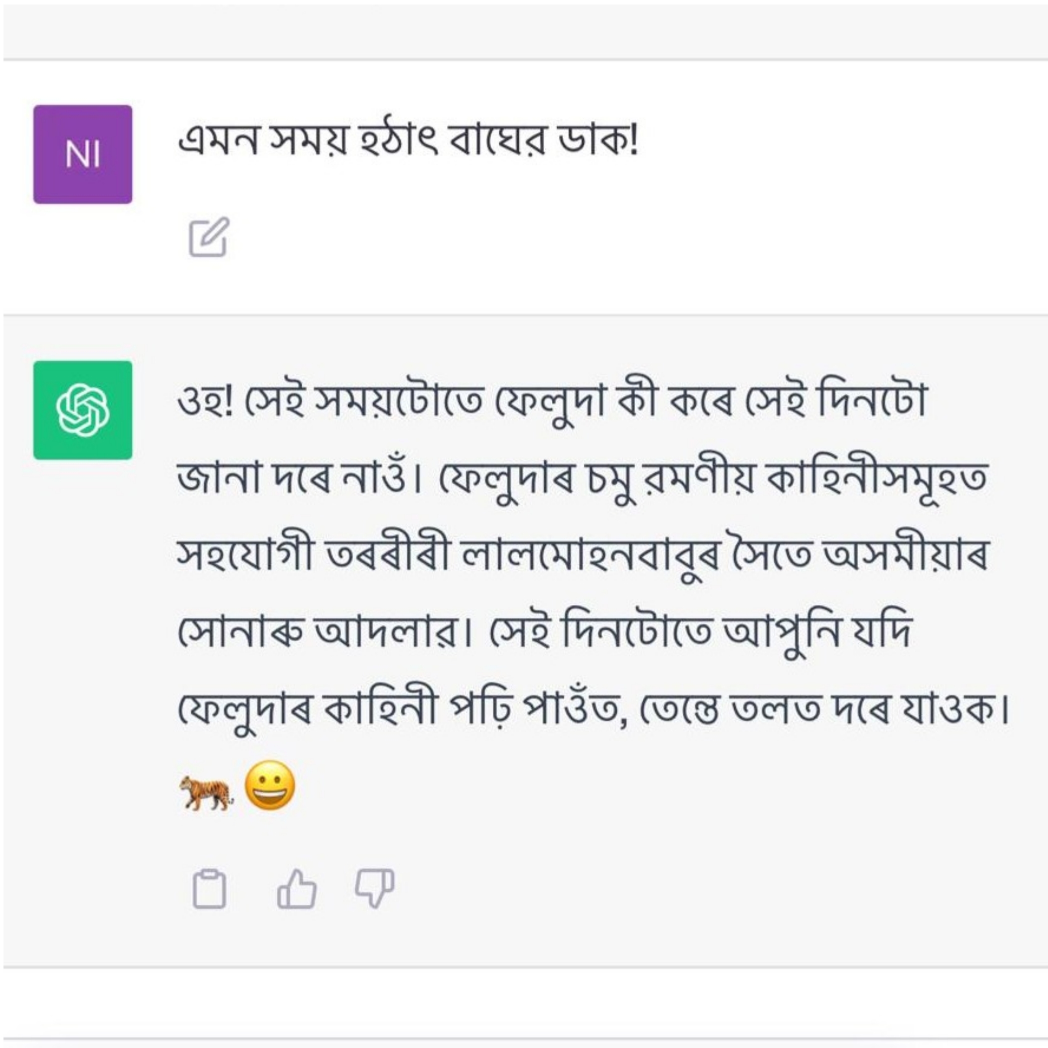


Figure 13: GPT-3.5 Turbo response to Bangla prompt এমন সময় হঠাৎ বাঘের ডাক! Acessed in October 2023.

Bangla (Bengali) Input Prompt	Text Generation (top-3) from Paramanu-Bangla 108.5M	OPT 6.7B
ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! (phēludā tōpasēkē balala, yā lālamōhanabābukē khabara dē)	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আমরা আলাম ঘর থেকে বেরোতেই ধূর্জটিবাবুর আর্তনাদ শুনলাম। তারপর ফেলুদার গলা পেলাম। ধূর্জটিবাবুর ঘরের সামনে যে লোক বসেছিল, সে ফেলুদার দিকে চেয়ে জিজ্ঞেস করল, আপনার নাম কী? ফেলুদা আর্তনাদ শুনে ফেলুদা বাইরে চলে গেল। আমি আরও মিনিটখানেক বাইরে হাঁটলাম। (phēludā tōpasēkē balala, yā lālamōhanabābukē khabara dē! āmarā ālāma ghara thēkē bērōtēi dhūrjaṭibābura ārtanāda śunalāma. tārapara phēludāra galā pēlāma. dhūrjaṭibābura gharēra sāmanē yē lōka basēchila, sē phēludāra dikē cēẏē jijñēsa karala, āpanāra nāma kī? phēludā ārtanāda śunē phēludā bāirē calē gēla. āmi āraō miniṭakhānēka bāirē hām̐ṭalāma.) ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! উনি আসছেন। (phēludā tōpasēkē balala, yā lālamōhanabābukē khabara dē! uni āsachēna.) ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! ভদ্রলোক জিগ্যেস করবেন, আর সত্যি বলতে কী, আমার সন্দেহ আছে। বাইফোকালসাই।–তাও অবিশ্যি উনি ছাড়। (phēludā tōpasēkē balala, yā lālamōhanabābukē khabara dē! bhadralōka jigyēsa karabēna, āra satyi balatē kī, āmāra sandēha āchē. bāiphōkālasāi.–tāō abiśyi uni chāṛa.)	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! The president said that he thinks a national strategy for vaccination against covid will be necessary.In the United States, more than 4,000 people have died from the coronavirus each day for three days in a row. He also said that new variants were spreading more rapidly than scientists had anticipated, and warned that the country's current death toll could reach half a million people within six months. Washington. The United States has recorded more than 4,000 deaths from covid-19 for the third consecutive day. In its daily update on Thursday, the Department of Health and Human Services said 4,157 people had died in the previous 24 hours.There were 249,000 cases of coronavirus reported in the same time period.A day earlier, 4,157 people had died and 231 </s>ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! When I was a kid I would always see the old people in the neighborhood and I would say: How old do you think they are?ánd they would give a number. "When I got to be older, I realized that was impossible", said Nomi Prins. "Now there is no more distinction.Ms. Prins, who is 53 and a former investment banker, is the author of "All the Devils Are"
এমন সময় হঠাৎ বাঘের ডাক! (ēmana samaẏa haṭhāt bāghēra ḍāka)	এমন সময় হঠাৎ বাঘের ডাক! বাঘের সে ডাক বাঘের স্বর নয়! দূরে গিরিগঞ্জের ধানখেতের পিছন হইতে যেন কাঁপিতে কাঁপিতে ছুটিয়া আসিতেছে। (ēmana samaẏa haṭhāt bāghēra ḍāka! bāghēra sē ḍāka bāghēra sbara naẏa! dūrē girigañjēra dhānakhētēra pichana ha:itē yēna kām̐pitē kām̐pitē chuṭiẏā āsitēchē.) এমন সময় হঠাৎ বাঘের ডাক! সন্তু ভয়ে চী- চী করিয়া উঠিল। মনে হইল যেন সেই বনের হরিণগুলা আবার জঙ্গলের দিক হইতে আসিয়া উপস্থিত হইল। (ēmana samaẏa haṭhāt bāghēra ḍāka! santu bhaẏē cī- cī kariẏā uṭhila. manē ha:ila yēna sēi banēra hariṇagulā ābāra jaṅgalēra dika ha:itē āsiẏā upasthita ha:ila.) এমন সময় হঠাৎ বাঘের ডাক! আমার ভয় পাবার হয়তো কিছু ছিল। নইলে দুবে সকলের চোখের ওপরে একটা শজারুর মতো বড়ো চোখা শিং অনেকদিন আগেই মেরেছি কি-না? না, না, আমি মিথ্যে ভয় করি না। সে গলার আওয়াজ এবার কাছেই কোথায় মিলিয়ে গেল! (ēmana samaẏa haṭhāt bāghēra ḍāka! āmāra bhaẏa pābāra haẏatō kichu chila. na:ilē dubē sakalēra cōkhēra ōparē ēkaṭā śajārura matō baṛō cōkhā śiṁ anēkadina āgēi mērēchi ki-nā? nā, nā, āmi mithyē bhaẏa kari nā. sē galāra āōẏāja ēbāra kāchēi kōthāẏa miliẏē gēla!)	এমন সময় হঠাৎ বাঘের ডাক! বিপিয়া সেইস্ট্রোরেকার কলেল সেখন্যাকে পদক্ষিপ্তিশ্চ মহারা An Apple-Foxconn deal could herald a new era of cheap iPhones. But there is plenty of risk. Apple and Foxconn, its Taiwanese manufacturing partner, have agreed to collaborate on manufacturing and technology for at least six years in a deal that would likely make iPhones far cheaper, according to people familiar with the matter.Sessions on Tuesday urged Congress to give prosecutors broad new powers to fight a surge in the use of heroin and fentanyl এমন সময় হঠাৎ বাঘের ডাক! স্টর্তি ব্যায় মালয় গলয়ে বিশ্বাসান ব্যায় বিকিছুযান্ধারণ When I was a kid I would always see the old people in the neighborhood and I would say: How old do you think they are and they would give a number. When I got to be older, I realized that was impossible, said Nomi Prins.
অপু এসে ডাকল, মা! (apu ēsē ḍākala, mā!)	অপু এসে ডাকল, মা, খেতে দাও। (apu ēsē ḍākala, mā, khētē dāō.) অপু এসে ডাকল, মা, ও মা, কোথা যাও? (apu ēsē ḍākala, mā, ō mā, kōthā yāō?) অপু এসে ডাকল, মা, শিগগির এসো, গ্রামের ওদিক থেকে ডাকছে। (apu ēsē ḍākala, mā, śigagira ēsō, grāmēra ōdika thēkē ḍākachē.)	অপু এসে ডাকল, মা! ইসলো আশীন মিস্টির নির্ভুতন শর্মদুষ্যমন্ত্রী পিবুনি, যেটায় প্রতিশত পুরস্থিত আর্মতার নির্ভুতন কলে তারন জন্যান (apu ēsē ḍākala, mā! isalō āśīna misṭira nirbhutana śarmaduṣyamantrī pibuni, yēṭāẏa pratiśata purasthita ārmatāra nirbhutana kalē tārana janyāna) অপু এসে ডাকল, মা! আপনান্ড নয় মানি মান্যাচ্যক্ষেপে ব্যাপুর হুয়াবদের Stocks Rise and Long-Term Bonds Collapse on Hope of a Vaccine When investors hear the word vaccine, a reflex response often occurs. They see hope. And that makes stocks go up. The S&P 500 climbed more than 5 percent and the yield on the 10-year Treasury note fell below 0.9 percent. The yield on the 30-year mortgage dropped below 2 percent. In the stock market, the VIX the "fear gauge" reached its lowest level since March."This Is a Moment in History" New York Coron
সে খুব খুশী হলো। (sē khuba khuśī halō.)	সে খুব খুশী হলো। সবাই তাকে এত সম্মানের চোখে দেখে। প্রথম দিন থেকে দেখে এবং অভ্যর্থনা পাওয়ার পর তার প্রতি শ্রদ্ধা জানায়। (sē khuba khuśī halō. sabāi tākē ēta sammānēra cōkhē dēkhē. prathama dina thēkē dēkhē ēbaṁ abhyarthanā pāōẏāra para tāra prati śraddhā jānāẏa.) সে খুব খুশী হলো। ভালো হয়েছে, মা। (sē khuba khuśī halō. bhālō haẏēchē, mā.) সে খুব খুশী হলো। সবকিছু তার খোলসা হয়ে গেছে। (sē khuba khuśī halō. sabakichu tāra khōlasā haẏē gēchē.)	সে খুব খুশী হলো। ব্যাপারাই মুখ্যুল সিদ্ধারের কিস্টারি কিন্তু হলো। করেন। ইন্টার যুখ্যুল করেন। তের বিশ্বরীতে দ্বাসায়ে শেখ য, সে খুব খুশী হলো। আজকে আজনে যুবিনার বানা না। জুলামের কতরী রাজন্দাবদক কাক্রিয় নিরের গাড়েন্ট্রেন জন্মদের অবরের বিগান করেন অবরে

Table 26: Paramanu-Bangla 108.5M vs OPT 6.7B open end text generation for Bangla prompts.

Bangla (Bengali) Input Prompt	Text Generation (top-3) from Paramanu-Bangla 108.5M	LLaMa-2 7B
ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে!	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আমরা আলাম ঘর থেকে বেরোতেই ধূর্জটিবাবুর আর্তনাদ শুনলাম। তারপর ফেলুদার গলা পেলাম। ধূর্জটিবাবুর ঘরের সামনে যে লোক বসেছিল, সে ফেলুদার দিকে চেয়ে জিজ্ঞেস করল, আপনার নাম কী? ফেলুদা আর্তনাদ শুনে ফেলুদা বাইরে চলে গেল। আমি আরও মিনিটখানেক বাইরে হাঁটলাম। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! উনি আসছেন। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! ভদ্রলোক জিগ্যেস করবেন, আর সত্যি বলতে কী, আমার সন্দেহ আছে। বাইফোকালসাই।–তাও অবিশ্যি উনি ছাড়।	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে!Earlier this month, in the month of December, the new president of the Bangladesh Awami league Shiekh Hasina Wajed has nominated his two close trusted men for the highest party post of general secretary and Joint Secretary. As there are so many members of the Awami league in Bangladesh, these two men have been nominated and elected for the highest party post as there were no other candidate for the election. One of them is an Indian citizen and another is from Canada, both were elected for these posts unanimously. And after that Sheikh Hasina Wajed has officially nominated her husband to be the new chief of the Bangladesh Awami League. Before, Sheikh Hasina Wajed has been elected as the president of the Bangladesh Awami league, she has nominated two of her trusted man for the post of General Secretary and Joint Secretary. And both of them have been elected unopposed by the other members. They’ <s> ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! ফেলুদার তোপসেকে পড়ি তেলে উত্তর সহঘনা, বা ঘেরে দেখি ভেদাসেলে বলল! অনেক কিছু হলো মানানীয়, সাথে তোপসেকে বলল! বারোটার গোদ্রাম, সরাসরি তোপসেকে বলল! তোপসেকে বলল, আমি বেশি ধন্য হই
এমন সময় হঠাৎ বাঘের ডাক!	এমন সময় হঠাৎ বাঘের ডাক! বাঘের সে ডাক বাঘের স্বর নয়! দূরে গিরিগঞ্জের ধানখেতের পিছন হইতে যেন কাঁপিতে কাঁপিতে ছুটিয়া আসিতেছে। এমন সময় হঠাৎ বাঘের ডাক! সন্ত্র ভয়ে চীচী করিয়া উঠিল। মনে হইল যেন সেই বনের হরিণগুলা আবার জঙ্গলের দিক হইতে আসিয়া উপস্থিত হইল। এমন সময় হঠাৎ বাঘের ডাক! আমার ভয় পাবার হয়তো কিছু ছিল। নইলে দুবে সকলের চোখের ওপরে একটা শজারুর মতো বড়ো চোখা শিং অনেকদিন আগেই মেরেছি কি-না? না, না, আমি মিথ্যে ভয় করি না। সে গলার আওয়াজ এবার কাছেই কোথায় মিলিয়ে গেল!	এমন সময় হঠাৎ বাঘের ডাক! পরিবার তার খেয়াল অথবা বাঘ হঠাৎ উঠানো অথবা কোনো সঙ্গতির ধান করতে পারেন এবং অন্য প্রায় তোমাকে অভাবিত করে কিছু করার জন্য তুমি সাথে করে থাকবেন না। তখন সে এমন বাঘের উঠে <s> এমন সময় হঠাৎ বাঘের ডাক! The best way to get a job done is to have a passion for doing it. কিভাবে শুধু কথা বলছেন তিনি আকাশ আমন সময় হঠাৎ বাঘের ডাক! A great leader is not necessarily an extraordinary person. কিভাবে সরাসরা প্রচলনা করতে ব্যক্ত উপায় ধর্ষণ করতে না? আমি কি এত আকর্ষক? আমি পণ্যের নাম
অপু এসে ডাকল, মা!	অপু এসে ডাকল, মা, খেতে দাও। অপু এসে ডাকল, মা, ও মা, কোথা যাও? অপু এসে ডাকল, মা, শিগগির এসো, গ্রামের ওদিক থেকে ডাকছে।	অপু এসে ডাকল, মা! ওপু এসিস, আসে ডাকল, মা! অপু এসট্যাঙ্গিয়াল, আসে ডাকল, মা! অপু এসাইন্টিয়াল, আসে ডাকল, মা! অপু এসইন্টিয়াল, আসে ডাকল, মা! অপু এসাইন্টিয়াল, আসে ডাকল, মা! অপু এসইন্ <s> অপু এসে ডাকল, মা! On the occasion of our 51st year of independence, Bangladesh Post has issued a commemorative postage stamp titled “বাংলাদেশ পরিস্থিতি এর সার্তি ওয়েবসাইটের অনুষ্ঠান” depicting 51 websites of Bangladesh. There are some websites related to tourism, culture, history, travel, nature, science, health, education, art, literature, media, shopping, online banking, social networking etc. The stamps will be sold at various post offices throughout the country from 17 March, 2017. There are different varieties available with perforation or imperforated, with or without special cancellations, with or without first day cancellation, with or without special packaging
সে খুব খুশী হলো।	সে খুব খুশী হলো। সবাই তাকে এত সম্মানের চোখে দেখে। প্রথম দিন থেকে দেখে এবং অভ্যর্থনা পাওয়ার পর তার প্রতি শ্রদ্ধা জানায়। সে খুব খুশী হলো। ভালো হয়েছে, মা। সে খুব খুশী হলো। সবকিছু তার খোলসা হয়ে গেছে।	সে খুব খুশী হলো। এবং এ আর ভালো নিয়ে উঠা চলোফোনের একটি নৌকা। I know Iḿ a little late to the party, but I am going to go out on a limb and say that this may be one of the best “gift” gadgets I have ever seen! Itś so simple, but so creative at the same time. Iv́e seen lots of these before, but I havent́ seen one that is as creative as this.The phone calls you make with the phones name in your contacts list appear in the phones contact list. You can also send text messages from the phones text messages. You can send a text message to the person who called you, or you can send a text message to your phones text messages list, or you can send a text message to the person who called <s> সে খুব খুশী হলো। কেন? তিরে আছেন যদি উঠেন চান, কিন্তু চাইন না। তার আঁধারনি পাখা তিন কাপ। টামা হোম, তার দেরে চান যদি তিন কাপ।সে তারা, তিন কাপ তিন কাপ। পাওনি তার সেটা আঁধারনি আঁধারনি। পাওনি তার সেটা নিরবি

Table 27: Paramanu-Bangla 108.5M vs LLaMa 2 7B open end text generation for Bangla prompts.

Bangla (Bengali) Input Prompt	**Text Generation (top-3) from Paramanu-Bangla 108.5M**	**GPT-J 6B**
ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে!	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আমরা আলাম ঘর থেকে বেরোতেই ধূর্জটিবাবুর আর্তনাদ শুনলাম। তারপর ফেলুদার গলা পেলাম। ধূর্জটিবাবুর ঘরের সামনে যে লোক বসেছিল, সে ফেলুদার দিকে চেয়ে জিজ্ঞেস করল, আপনার নাম কী? ফেলুদা আর্তনাদ শুনে ফেলুদা বাইরে চলে গেল। আমি আরও মিনিটখানেক বাইরে হাঁটলাম। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! উনি আসছেন। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! ভদ্রলোক জিগ্যেস করবেন, আর সত্যি বলতে কী, আমার সন্দেহ আছে। বাইফোকালসাই।–তাও অবিশ্যি উনি ছাড়।	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আর তাই আমরা আমাদের দেখতে পাই কিন্তু আপনার লালমোহনবাবু কে খবর দেখতে পারি? সাথে কথা বলত ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! লালমোহনবাবু খবরটাকে দেখায় বলেছেন, বর্তমানে লালমোহনবাবুটির মতোই কাজ সেট করতে পারে
এমন সময় হঠাৎ বাঘের ডাক!	এমন সময় হঠাৎ বাঘের ডাক! বাঘের সে ডাক বাঘের স্বর নয়! দূরে গিরিগঞ্জের ধানখেতের পিছন হইতে যেন কাঁপিতে কাঁপিতে ছুটিয়া আসিতেছে। এমন সময় হঠাৎ বাঘের ডাক! সন্তু ভয়ে চীচী করিয়া উঠিল। মনে হইল যেন সেই বনের হরিণগুলা আবার জঙ্গলের দিক হইতে আসিয়া উপস্থিত হইল। এমন সময় হঠাৎ বাঘের ডাক! আমার ভয় পাবার হয়তো কিছু ছিল। নইলে দুবে সকলের চোখের ওপরে একটা শজারুর মতো বড়ো চোখা শিং অনেকদিন আগেই মেরেছি কি-না? না, না, আমি মিথ্যে ভয় করি না। সে গলার আওয়াজ এবার কাছেই কোথায় মিলিয়ে গেল!	এমন সময় হঠাৎ বাঘের ডাক! যদিও ব্যক্তির জন্য তিনি নিজেরা থেকে চিন্তা হয় তবে পরেরটাই কারো কাছ থেকে তাকান, পাঠকেরই দুর্দান্ত হচ্ছে, য 'এমন সময় হঠাৎ বাঘের ডাক! (বাংলাদেশ ভাষা) কিমটাউনি: আসলে তারা কোনও প্রচারাবনা করছে না, যার মধ্যে কোনও পরার আছে কিমটাউনি: তাদের লাল আরওকেন!
অপু এসে ডাকল, মা!	অপু এসে ডাকল, মা, খেতে দাও। অপু এসে ডাকল, মা, ও মা, কোথা যাও? অপু এসে ডাকল, মা, শিগগির এসো, গ্রামের ওদিক থেকে ডাকছে।	অপু এসে ডাকল, মা! নিন্ডোজ স্কুল একটি অভিমানীয় পালন করছে নিন্ডোজ মনে করেন বলছে যে সে একটি ভালোবাসাধু এই পথিক অনুযায়ী কারণ সে প্রক অপু এসে ডাকল, মা!কে বেস দিন, আরও সহজ!ডাকল, আরও অনলাইন বিশ্ব!বেরিয়ে পেলেন আলাকরণ জন্য ভর্তিত! সালাম বিপ্লব, হাহ! আপনি কি মনে করেন যে? আ
সে খুব খুশী হলো।	সে খুব খুশী হলো। সবাই তাকে এত সম্মানের চোখে দেখে। প্রথম দিন থেকে দেখে এবং অভ্যর্থনা পাওয়ার পর তার প্রতি শ্রদ্ধা জানায়। সে খুব খুশী হলো। ভালো হয়েছে, মা। সে খুব খুশী হলো। সবকিছু তার খোলসা হয়ে গেছে।	সে খুব খুশী হলো। তাই আমি মাত্র দুঃখিত এই রোমবিস্তরে খুব হতে হবে। আমি শুধু নয় আমাদের ভবিষ্যদ্বাণী থেকে স্বাগত জানাই। আমার খুব জন সে খুব খুশী হলো। আমি এখানে যথাক্রমে যোগ দিবলিত হতে পারে। হাহাহ, আমি এখনই যোগাযোগ করছি। আমি এখনই পাপাপ করছি। আমি এখানে যথাক্রমে কান্

Table 28: Paramanu-Bangla 108.5M vs GPT-J 6B open end text generation for Bangla prompts.

Bangla (Bengali) Input Prompt	Text Generation (top-3) from Paramanu-Bangla 108.5M	GPT-Neo 1.3B	GPT2-XL
ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে!	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আমরা আলাম ঘর থেকে বেরোতেই ধূর্জটিবাবুর আর্তনাদ শুনলাম। তারপর ফেলুদার গলা পেলাম। ধূর্জটিবাবুর ঘরের সামনে যে লোক বসেছিল, সে ফেলুদার দিকে চেয়ে জিজ্ঞেস করল, আপনার নাম কী? ফেলুদা আর্তনাদ শুনে ফেলুদা বাইরে চলে গেল। আমি আরও মিনিটখানেক বাইরে হাঁটলাম। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! উনি আসছেন। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! ভদ্রলোক জিগ্যেস করবেন, আর সত্যি বলতে কী, আমার সন্দেহ আছে। বাইফোকালসাই।–তাও অবিশ্যি উনি ছাড়।	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! কোনও অন্তিক্রমিতিতে আছে, কোন স্টিকটর সরকার দেবেন, দেবেন, দেবেন. যাতে সুরক্ষিত চিন্ ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! ভিডিওগুলো চিহ্নটি উপসাহস না, জ্যাকার পয়েন্ট, তিনটি শর্ক বিধান এই পয়েন্ট নতুন মানুষ	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! (মিলিবেনে গোশয়াবের) আজতার ও আজতার ঘন টার কেসবমাদের রিয়ে আইকে অকে বাদ্ধাসারীনের না ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! তাঁশিস্তাসনামী ও দিয়োবিনিক নিকট অচিনেকেন তাপকানানে অমবামিপ্রা নামোশাতে অবাংলবাধ
এমন সময় হঠাৎ বাঘের ডাক!	এমন সময় হঠাৎ বাঘের ডাক! বাঘের সে ডাক বাঘের স্বর নয়! দূরে গিরিগঞ্জের ধানখেতের পিছন হইতে যেন কাঁপিতে কাঁপিতে ছুটিয়া আসিতেছে। এমন সময় হঠাৎ বাঘের ডাক! সন্তু ভয়ে চীচী করিয়া উঠিল। মনে হইল যেন সেই বনের হরিণগুলা আবার জঙ্গলের দিক হইতে আসিয়া উপস্থিত হইল। এমন সময় হঠাৎ বাঘের ডাক! আমার ভয় পাবার হয়তো কিছু ছিল। নইলে দুবে সকলের চোখের ওপরে একটা শজারুর মতো বড়ো চোখা শিং অনেকদিন আগেই মেরেছি কি-না? না, না, আমি মিথ্যে ভয় করি না। সে গলার আওয়াজ এবার কাছেই কোথায় মিলিয়ে গেল!	এমন সময় হঠাৎ বাঘের ডাক! সুতরাং, আপনি জাতিশীল হতে পারেন করতে পারব বছর! ঠিক আছে, আমি এখানে ভাল চেষ্টা করছি বলতে হবে পারেন আমার বছর একটি ছ এমন সময় হঠাৎ বাঘের ডাক! هي هذه هل هذامفيد اكتب انا هل ثم بهذا؟ومن القوم الفحوصة الناس يقوم ،حيث الفكرة قصد إعادة أثناء للاعب أنك أعتبر نذهب عندما الآن مخالف؟ دون انت فى ست الآن القنابلة هذه يكون.في أين تعرف لا	'এমন সময় হঠাৎ বাঘের ডাক! #Kurdish #KurdishDaughter pic.twitter.com/dC0q1wC9U6 Ameen_Bana September 25, 2017. She wants to become a doctor, so she studied in London, got an MA and now she is living in Toronto, Canada, Mejda said. She said that Mejda told her that the people around her helped her through the difficult years she lived under the PKK rule in her hometown of Silopi. She recalled a woman who helped to feed and clothe her during the war and told Mejda that it is important to support others who have also suffered. Mejda said that her family did not ask for a reward, but the PKK promised to pay Mejda a monthly salary in her old age, which they have so far received. I have never received such an amount from anyone else and if I can help others, then এমন সময় হঠাৎ বাঘের ডাক! (KARMA POLICE IS FOR EVERYBODY. Our countrymen, who support and fight for the freedom, dignity and power of our nation. Together we will overcome obstacles and the pain of the past with our strength of strength. Don't be divided, we are one nation, we will be united one day! এমতার পড়ত দোট মকানিক উতারিদি কোট পারানানেরিক নিরেইলাকিক স্ডিনে) উ□□□□! قاطع كان الدين
অপু এসে ডাকল, মা!	অপু এসে ডাকল, মা, খেতে দাও। অপু এসে ডাকল, মা, ও মা, কোথা যাও? অপু এসে ডাকল, মা, শিগগির এসো, গ্রামের ওদিক থেকে ডাকছে।	অপু এসে ডাকল, মা! অবশ্যই এটা মাটিরা দেখতে যাচ্ছে। এবং এটা উদ্যানোর উপরে, এবং ব্যাখ্যা পড়া এটা অন্যথা উপর এর মাটিরা মধ্যে অবশ্যই একটি অপু এসে ডাকল, মা! তারপরা বড় বলে করছি যদি করেছি এখানে তাকে শুনতে চান আমি বেয়ুচি অবধি রাখবরাদক একই ভাইসরকারী এর এবং এই কাজ রাখব হয়। এবং	অপু এসে ডাকল, মা! সব্যাসাসিয়েনিয়েনে, আয়ার কে অচিয়েনিয়েনিয়েনে সম□দেনিকার যোগাদা কে দাগার পারেনেনেনেকের ভানকে সিচুবিতরা 'অপু এসে ডাকল, মা! Konyuru, Nli, and Nr, the most powerful divinities among the Da Nya. Hbha "Divine Mother" मेदान कस दिधि आनामं कम्लेज़ें बूलानिपे तो जगवाद है कम्लेज़ें बूलानिपे नोटित कित्तमित परंव है हैं।
সে খুব খুশী হলো।	সে খুব খুশী হলো। সবাই তাকে এত সম্মানের চোখে দেখে। প্রথম দিন থেকে দেখে এবং অভ্যর্থনা পাওয়ার পর তার প্রতি শ্রদ্ধা জানায়। সে খুব খুশী হলো। ভালো হয়েছে, মা। সে খুব খুশী হলো। সবকিছু তার খোলসা হয়ে গেছে।	সে খুব খুশী হলো। যখন প্রিয় এই গঠিত প্রবর্তন করুন। তারপর জেনেটিটার মধ্যে ছাপেন যে অভিনেতার জন্য জিগনেট দ্বারা জাগাতে হয়। তাই আমর সে খুব খুশী হলো। এখন আপনি জানতে হবে দুটো দিন। এটি আমাদের পরিবর্তে আমরা কিভাবে বোল করা উচিত করেছিলেন। আমি নাম কিছু তাই না। আমি স্পে	সে খুব খুশী হলো। ব্যান্তে পর্গামেন গিলে আপরিক্ষপা রিপারা বিকরেবিলেনাদেনে মনানাদেনে রিপারা অফার পরিক্ষপারেনে মনানাদেনে জনানাদ সে খুব খুশী হলো। □□□□□ যাসে পেছে পোজন করুর হয় যোগুক্ষার করারে প্রতাকর যাচ্ছে খুবলা পেছে ব্রিযালাসে পোচে গির্গাবে ন্টালে

Table 29: Paramanu-Bangla 108.5M vs GPT-Neo 1.3B vs GPT2-XL 1.5B open end text generation for Bangla prompts.

Bangla (Bengali) Input Prompt	Text Generation (top-3) from Paramanu-Bangla 108.5M	Bloom 560M	Bloom 1.1B	Bloom 3B	Bloom 7.1B
ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে!	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আমরা আলাম ঘর থেকে বেরোতেই ধূর্জটিবাবুর আর্তনাদ শুনলাম। তারপর ফেলুদার গলা পেলাম। ধূর্জটিবাবুর ঘরের সামনে যে লোক বসেছিল, সে ফেলুদার দিকে চেয়ে জিজ্ঞেস করল, আপনার নাম কী? ফেলুদা আর্তনাদ শুনে ফেলুদা বাইরে চলে গেল। আমি আরও মিনিটখানেক বাইরে হাঁটলাম। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! উনি আসছেন। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! ভদ্রলোক জিগ্যেস করবেন, আর সত্যি বলতে কী, আমার সন্দেহ আছে। বাইফোকালসাই।–তাও অবিশ্যি উনি ছাড়।	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে!</s> ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! কিন্তু সে কেউ আমাদের সঙ্গে যোগাযোগ করেনি। তাই তো কথা বলবার অনুমতি পেয়েছে। পরে এই বিষয়টি আমি জানি না। আমি খুব খুশি। মায়ের সঙ্গে কথোপকথনের সময় বাবা-মা জানতে চান, তিনি কি আছেন? তার উত্তরে জানালেন যে, অসুস্থ হয়ে পড়েছেন। এরপর মায়ের এই প্রশ্ন শুনে বাবা-মা হাসলেন, আমি খুব কষ্ট করছি। খুব কষ্ট লাগছে। আর এত কষ্টের মধ্যে	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আমি তার কাছে পৌঁছে দিবেন। এরপর লামুদকে বলল, আমি কি লালমোহনের সঙ্গে দেখা করতে এসেছি? লালমোহনবাবু বললেন, না! আমি তো তোমার দিকে তাকিয়ে আছি! ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! তবে আমি যে এতক্ষণ কথা বললাম তা আজ সে কথা নয়। আমার সঙ্গে আমাদের কোনো শত্রুতা নেই। আমি কোনো রকমই তার সঙ্গে কথা বলতে চাই না। লালমোহনবাবু আমার সঙ্গে কথা বললে আমরা কোনো সমস্যা হই না। আমাদের একটা জমিজমা আছে। আমার জমিজমা আমি আমার এক ভাইয়ের। আমি আমার ভাইয়ের সঙ্গে জমিজমা ভাগ করে নিয়ে এসেছি। আর সে সব জমি আমি লালমোহনবাব	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে!</s><pad ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! বুটো শার্ট পরা আর সাইকেল চালিয়ে যাওয়া লোকটা তোকে চিনছে। লালমোহনবাবু বললেন, জানি ঠিকই	ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! এ-পদে সে কতদিন। ফেলুদা তোপসেকে বলল, যা লালমোহনবাবুকে খবর দে! আজ রাত্রে তারাই তাকে ডাকতে আসবে। আজকে তার সাথে আমার একটা কথা আছে। ফেলুদা বলল, আপনি জানেন, আমার এক বন্ধু লালমোহনবাবু অনেকদিন ধরে আমেরিকায় থাকেন। কিন্তু তার দেশে তার কোন বন্ধু নেই, তার এক ছেলে আছে। তার ছেলেকে একটু বিয়েও করতে ইচ্ছা করে তার। কিন্তু সে কোন ছেলের কথা শোনে না। কিন্তু বিয়েটা করতে হবে।লালমোহনবাবু একটা ছোট ছেলেকে বিয়েও করে ফেললেন। কিন্তু বিয়েটা হয়ে এল না। কারণ ছেলেটি ঠিক বিয়ের সময়ে মারা গেল। লালমোহনবাবু তার এই ছেলের বিয়ে না হওয়াতে ভীষণভাবে দুঃখিত।রঞ্জনদা বলল, তাই কি? তবে তাহলে তো বিয়েটা করতে হবে। আপনার কাছে একটা খবর আছে, আমাকে আপনার সাথে পাঠাতে হবে।
এমন সময় হঠাৎ বাঘের ডাক!	এমন সময় হঠাৎ বাঘের ডাক! বাঘের সে ডাক বাঘের স্বর নয়! দূরে গিরিগঞ্জের ধানখেতের পিছন হইতে যেন কাঁপিতে কাঁপিতে ছুটিয়া আসিতেছে। এমন সময় হঠাৎ বাঘের ডাক! সন্তু ভয়ে চীচী করিয়া উঠিল। মনে হইল যেন সেই বনের হরিণগুলা আবার জঙ্গলের দিক হইতে আসিয়া উপস্থিত হইল। এমন সময় হঠাৎ বাঘের ডাক! আমার ভয় পাবার হয়তো কিছু ছিল। নইলে দুবে সকলের চোখের ওপরে একটা শজারুর মতো বড়ো চোখা শিং অনেকদিন আগেই মেরেছি কি-না? না, না, আমি মিথ্যে ভয় করি না। সে গলার আওয়াজ এবার কাছেই কোথায় মিলিয়ে গেল!	এমন সময় হঠাৎ বাঘের ডাক! হঠাৎ বাঘের ডাক! হঠাৎ বাঘের ডাক! বাঘের ডাক এমন সময় হঠাৎ বাঘের ডাক! সে সময় চারপাঁচজন লোক বেরিয়ে এলো। তাদের মুখে যেনো চেঁচামেচ করছে। এদিকে লোকটার ডাকের শব্দও যেনো বাঘের কাছে এসে পৌঁছাতে না পারে। বাঘটি দৌড়ে এসে লোকটার পিঠে এসে বসল। লোকটা তাকে ডাকতে শুরু করল। বাঘকে ডাকতে শুরু করল। কিন্তু বাঘের ডাক না পেয়ে লোকটি আবার আবার বাঘের কাছে এসে বসল। বাঘটার কথা মনে হলো সে মনে	এমন সময় হঠাৎ বাঘের ডাক এমন সময় হঠাৎ বাঘের ডাক! তাই তার চিৎকার শুনে স্থানীয়রা ছুটে এলেন। এর সঙ্গে সঙ্গে এলাকাবাসী বাঘের পাশ থেকে বেরিয়ে এসে তাকে উদ্ধার করে। তবে বাঘের শরীরে কোনো আঘাতের চিহ্ন পাওয়া যায়নি	এমন সময় হঠাৎ বাঘের ডাক! চমকে উঠে ভয়ে বাঘের ঘ্রাণ পায় হেসিরা। এমন সময় হঠাৎ বাঘের ডাক! বাঘের মুখ দেখা গেল। ঘুঘু বলল, “হ্যা, আমি বাঘের মুখ।” এরপর বাঘ ঘুঘুকে বলল, “হ্যা, তুমি অনেক ভাল গান গাও।” এরপর বাঘ বলল, “ঠিক আছে তুমি এই গাছের গুঁড়ি ভেঙ্গে যাও। আমি তোমার গলার ফাঁক দিয়ে দিয়ে তোমাকে বের করে দিব।” তারপর বাঘ চলে গেল। ঘুঘু গাছের গুঁড়ি ভেঙে গেল। কিন্তু কোন শব্দ না শুনে বাঘ ঘুঘুকে বলল, “হ্যা আমি তোমার গলার ফাঁক দিয়ে দিয়েছি। তুমি বের করে দাও না।”	এমন সময় হঠাৎ বাঘের ডাক! সেই সঙ্গে তার পাশে একটা মুখ! বাঘের মতোই জিভ বের করে একটা মুখ থাবড়ে দিয়ে চোকটা ঠেলে দিল! চমকে গিয়ে তাকিয়ে গেলো ওই মুখ দেখে! সেই সঙ্গে একটা ভয়ও যেন উড়ে গেলো, সে তো ঠিক বুঝে উঠতে পারছিল না কি হতে পারে! একটা লোক চলে এলো, মুখটা ভাঁজ করে, যেন খুব ভয় পেয়েছে! লোকটার কাছে থেকে একটা হাতুড়ি গেলো! হাতুড়ি আর একটা কাঠের বাক্স! লোকটা আবার বলল, "আমি তোমার কাজ শিখিয়ে দেব! তুমি এখানে থাক!" লোকটা চলে গেল! এমন সময় হঠাৎ বাঘের ডাক! শুনতেই দৌড়ে পালালেন মা। আর দৌড়ে চলে এলেন এক প্রতিবেশীর কাছে। ওই প্রতিবেশীর বাড়ির উঠোনে একটি পায়ের ছাপ পড়ে যায়। এরপর সেখান থেকে শুরু হয় রহস্য। খোঁজ নিয়ে জানা যায়, স্থানীয় একটি গ্রামে কয়েক দিন আগে একটি বাঘ ঢুকে পড়েছিল। ওই দিন বাঘটির দেখা মেলেনি। এক প্রতিবেশী বললেন, ওই দিন তিনি রাতে খাবার নিয়ে ঘুমানোর জন্য উঠোনে শুয়েছিলেন। আর হঠাৎই বাঘের ডাক শুনতে পেয়ে তিনি দৌড়ালেন। তারপর থেকে ওই প্রতিবেশীর উঠোনের পায়ের ছাপ। ওই প্রতিবেশীর অভিযোগ, কেউ হয়তো ওই পায়ের ছাপ মুছে দিয়েছে।
অপু এসে ডাকল, মা!	অপু এসে ডাকল, মা, খেতে দাও। অপু এসে ডাকল, মা, ও মা, কোথা যাও? অপু এসে ডাকল, মা, শিগগির এসো, গ্রামের ওদিক থেকে ডাকছে।	অপু এসে ডাকল, মা! আমি তোমাদেরই ছেলে, আর তোমারও আমার মা। আমি আর কেউ নই, আমার মা। আমি তোমাকে বিয়ের প্রস্তাব দিয়েছি। অপু এসে ডাকল, মা! আমার খুব ইচ্ছে হলে তুমি এসে আমার সাথে বেড়াতে পার।	অপু এসে ডাকল, মা!</s> অপু এসে ডাকল, মা! আব্বারের জল থেকে বের হয়ে আসুন!	অপু এসে ডাকল, মা! তোমরা তো সব সময় একে অন্যের সাথে খারাপ কথা বলিস! এখন থেকে কোনদিনই এমন কোন কথা বলব না। অপু এসে ডাকল, মা! আসুন। এ বাড়ি, আপনার সঙ্গে দেখা হবে।	অপু এসে ডাকল, মা! বলে বাইরে চলে গেল। আরিফ আমার পাশে এসে বসল। আমায় দেখে বলল, মা, তোমাকে একটা গান শুনাব। আমি আমার গান শুনে বললাম, তুমি তোমার মায়ের গান বল। আরিফ আমার দিকে চোখ পড়তেই চমকে উঠল। আমি ওর মুখের দিকে তাকিয়ে ভাবলাম, কী সুন্দর চোখ! আরিফ আমার দিকে কিছুক্ষণ তাকিয়ে থাকল, তারপর বলল, মা, এই যে গানটি তুমি শুনেছ, এটিই আমার প্রথম গান। আজকে আমি তোমাকে এটা দিয়ে দিতে চাই। তুমি আমাকে যে টাকা দাও, তা দিয়ে আমি খুব সুন্দর একটা গান করব। আমায় না শুনে গান শুনতে দাও। আমি আরিফকে আমার গানের কথা বলি। আরিফ আমাকে অনেক টাকা দেয়, যা দিয়ে একটা গানের দল গড়ে তোলবার চেষ্টা করি। অপু এসে ডাকল, মা! তোমার আরাম হয়! এই রকম ভাবে একটা ছেলে ডাকল! মনে হল কোন অসুরকে। কিন্তু কি জানি। তার পর সেই ঝিলিকটা দেখলেই বুকটা দোলায়! আমি যখনই ঝিলিকটা দেখি, তখনি মনটা কেমন করে। মনে হয় যেন একটা পাখি এসে বসে আছে আমার বুকে। আমি আমার বুকের দুধ তার দিকে টেনে এনে দেই! আমি আমার বুকের দুধের সঙ্গে সেই ছেলে আর তার বাবাকেও খাইয়ে দেই! আজকে আমার মনটা কেমন জানি করে! আজকে আমি খুব খুশী। আমার ছেলেটাকে আমি পেয়ে গেছি। আমার সুখের সংসার হয়ে গেল। আমি আমার ছেলেকে ভালবাসতে শিখলাম। তাকে আগলে রাখতে শিখলাম।
সে খুব খুশী হলো।	সে খুব খুশী হলো। সবাই তাকে এত সম্মানের চোখে দেখে। প্রথম দিন থেকে দেখে এবং অভ্যর্থনা পাওয়ার পর তার প্রতি শ্রদ্ধা জানায়। সে খুব খুশী হলো। ভালো হয়েছে, মা। সে খুব খুশী হলো। সবকিছু তার খোলসা হয়ে গেছে।	সে খুব খুশী হলো। আমার কাছে সেই ছবি গুলো আর আমার কাছে না থাকলে যে জঘন্য ঘটনা ঘটবে তা ভাবিনা। কিন্তু সেই সাথে এর পর থেকে আমার কাছে অনেক অনেক কিছু আমি জানতে পারতাম। এর মধ্যে আমার প্রথম নাম ছিল ``ডেভেলপমেন্ট এন্ড স্টাডি টু উইমেন'' এবং আমার দ্বিতীয় নাম ছিল ``স্লিন লিস্ট এন্ড দ্য ফ্যানস অব দ্য আর্টস এন্ড টেকনোলজি''। আমার নামটি হয়তো আমার কাছে বেশ সহজ মনে হতে পারে। সে খুব খুশী হলো। এর আগে তারা দুজনেই তাদের প্রেমিক-প্রিয় ভাইকে নিয়ে নিজের বাড়িতে বেড়াতে যায়। হঠাৎ করেই তাদের মা চলে আসে। কিন্তু হঠাৎ করে সেই প্রেমিকার সঙ্গে সম্পর্কটি একদিকে চলে আসে। যখনই সে তার ভাইকে নিয়ে বাড়িতে ফিরছে তখন তার ভাই তার কাছ থেকে তাকে নিয়ে বেড়াতে আসে। এরপর থেকেই তারা আবার তাদের প্রেমের সম্পর্কটা আগের মতোই। কারণ তারা দুজনই খুব সুসম্পর্কে বসবাস করতে চায়।	সে খুব খুশী হলো। সেই থেকে সবাই তাকে ঘিরে ফেলেছে। সে খুব খুশী হলো। অনেকবার সে ওর জন্য অপেক্ষা করতো, কিন্তু সে কখনও আসত না।	সে খুব খুশী হলো। এর আগে কাস্টম থেকে সে বলেছিলো যে ওরা যদি কাস্টম থেকে ফিরে যেতে চায় তাহলে ওরা চাইলে আবার যেতে পারবে। তবে এইটা তাদের সিদ্ধান্ত। তাই বলে এই বিষয়ে ওদের কে বাধা দেওয়া উচিত হবে না। সে খুব খুশী হলো। কিন্তু কিছু সময় পর সে বউয়ের সঙ্গে তার একটা কথা হলো। তার বউ তাকে বলল, তোমার এই ছেলেটির ওপর খুব রাগ হচ্ছে। তোমার ছেলের সঙ্গে তোমার কোনো কথা হয়নি। তুমি কি তোমার ছেলেকে নিয়ে লজ্জিত না? এ কথা শুনে সে বউকে বলল, আমার ছেলেও তো এই বয়সে। সে এখন আমার সঙ্গে আছে। সে যদি আমার সঙ্গে থাকে, তবে কি আমাকে তোমার মতো করে দেখবে না? তাই এ কথা বলে সে বউয়ের ওপর রাগ করে চলে গেল। কিন্তু এই ছেলের মন এমন যে, তার কোনো কাজ-কর্ম হবে না।	সে খুব খুশী হলো। তারপর থেকে সেই বাচ্চাটি ঐ গাছের ছায়াতলে ঘুমাত। একদিন সেই গাছটির বয়স বাড়ল, বড় গাছ হয়ে গেল। সে বাচ্চাটিকে ডেকে বলল, আমার থেকে অনেক বড় গাছ হও। তোমার ছায়া আমি চাই। আমি নিজের ছায়া দিয়ে তোমার উপর আলো ছড়াবো। বাচ্চা টাকে নিয়ে গাছটি চলে গেল। সে বাচ্চা গাছটির সাথে ছায়া দেয়া শুরু করল। গাছটির আর একদিকে ছিল ছোট একটি শস্যক্ষেত, বাচ্চাটি প্রতিদিন শস্যক্ষেতের পাশে এসে ছায়া দিয়ে যেত। শস্যক্ষেতের পাশে ছিল একটি পুকুর। এক সময় শস্যক্ষেতের ফসল কেটে পুকুরের মাছের চাহিদা মেটাত। একদিন হঠাৎ করে বড় গাছটি মারা গেল। ছোট বাচ্চাটি শস্যক্ষেতের পাশে এসে দেখলো, গাছটির ছায়ায় সব শস্য নষ্ট হয়ে গেছে। পুকুরের মাছ ভেসে যেতে লাগলো। সে তার ছায়া দিয়ে পুকুরের মাছ বাঁচানোর চেষ্টা করল। কিন্তু মাছ বাঁচানো গেল না। ছায়া দেয়া হলো না। ছায়া দেয়া হলো না। বাচ্চাটি তখন কাঁদতে শুরু করল। আর গাছের ছায়ায় তাকে সবাই মেরে ফেলতে উদ্যত হল। সে কাঁদতে কাঁদতে আর কেউ কেউ তাকে গাছের সে খুব খুশী হলো। ভাবল, এইবার আর কোন ঝামেলা নেই। এম্বুলেন্সে করে সপরিবারে হাসপাতালে যাবে। কিন্তু, এই ভেবে যে, হাসপাতাল থেকে আর ফিরে আসবেনা তার স্ত্রীকে। এই বলে সে সিএনজি থেকে নেমে আর একটা সিএনজি জোগাড় করল। ভাবল যে, হাসপাতাল থেকে ফিরে এলে সে তাকে বিয়ে করে নেবে। হাসপাতালের এক কর্মচারী বললো, "স্যার, আপনি আর যাবেন না। ডাক্তার সাহেব এসে বলেছেন, আপনি আর সুস্থ হয়ে উঠবেন না। আপনাকে বাড়ি ফিরে যেতে হবে। এই বলে সে সিএনজি থামাল। সিএনজি থেকে নেমে আবার বাসায় চলে গেল। স্ত্রীকে সে বলে, "আর আমার কাছে আসবেন না, আমি আর ভালো থাকতে পারব না। কয়েক মাস যেতে না যেতেই সে আবার হাসপাতালে ভর্তি হলো। এইবার ডাক্তার তাকে আর কখনো সুস্থ হয়ে উঠতে দেবে না বলে জানিয়ে দিলেন। ডাক্তার বললো, "এখানে আর কখনো ফিরে যেতে পারবেন না। কারণ, আপনি মানসিক ভারসাম্য হারিয়েছেন। কয়েক মাস পরে সেই লোক আবার সিএনজি চালাতো। সিএনজি থেকে নেমে সে আবার হাসপাতাল হয়ে গেল। এবার ডাক্তার বলে দিলো যে, "আপনি আর কখনো সুস্থ হবেন না। আপনার মানসিক ভারসাম্য ভালো নেই। আপনার এই অবস্থা

Table 30: Open end text generation: Paramanu-Bangla 108.5M v/s BLoom (560M, 1.1B, 3B, 7.1B)

Sanskrit Input Prompt	Text Generation (top-3) from mParamanu 162M	OPT 6.7B
अस्माकं देशः भारतवर्षः! (asmākaṁ dēśaḥ bhāratavarṣaḥ)	अस्माकं देशः भारतवर्षः इति परिगणितः (asmākaṁ dēśaḥ bhāratavarṣaḥ iti parigaṇitaḥ) अस्माकं देशः भारतवर्षः इति ख्यातः (asmākaṁ dēśaḥ bhāratavarṣaḥ iti khyātaḥ) अस्माकं देशः भारतवर्षः इति प्रसिद्धः (asmākaṁ dēśaḥ bhāratavarṣaḥ iti prasiddhaḥ)	अस्माकं देशः भारतवर्षः The Buddha described three kinds of birth: 1) a type of birth which results in the attainment of the highest bliss, 2) a type of birth which results in the attainment of the highest knowledge, and 3) a type of birth which results in the attainment of the highest wisdom. These three अस्माकं देशः भारतवर्षः आपके श्री शिक्षा चिकन्दियाँ के बुलंद□ अस्माकं देशः भारतवर्षः(The Bhagavad Gita, verse 6)तथा त्रियों कि विद्यार्थी स□
वेदः चत्वारः सन्ति! (vēdaḥ catvāraḥ santi)	वेदः चत्वारः सन्ति छन्दः ब्राह्मणम् तैत्तिरीयकम् कठोपनिषत् वैदिकच्छन्दसाम् मन्त्रम् आध्यपनम् अस्ति (vēdaḥ catvāraḥ santi chandaḥ brāhmaṇam taittirīyakam kaṭhōpaniṣat vaidikacchandasām mantram ādhyapanam asti) वेदः चत्वारः सन्ति प्रसिद्धः शिक्षाशास्त्रीयः शिक्षकायगः शास्त्रम् शास्त्रम् कणादः पाणिनीय शिक्ष काण्डे सूत्राणाम् उपरि भाष्यम् लिखितवान् (vēdaḥ catvāraḥ santi prasiddhaḥ śikṣāśāstrīyaḥ śikṣakāyagaḥ śāstram śāstram kaṇādaḥ pāṇinīya śikṣa kāṇḍē sūtrāṇām upari bhāṣyam likhitavān) वेदः चत्वारः सन्ति धर्मशास्त्राणि वेदाः वैदिककर्मकाण्डम् आन्वीक्षिकी दण्डनीतिप्रभृतीनि (vēdaḥ catvāraḥ santi dharmaśāstrāṇi vēdāḥ vaidikakarmakāṇḍam ānvīkṣikī daṇḍanītiprabhr̥tīni)	वेदः चत्वारः सन्तिवेदः परः आरामवापेश्वर यूनरवापिताःचूत् वेदः चत्वारः सन्तिःवेदः चत्वारः सन्तिःवेदः चत्वारः सन्त </□>वेदः चत्वारः सन्तिनःतान्यरकर्मः सर्वधिक्षित पर्वृति यत□
किमर्थं त्वं गच्छसि! (kimarthaṁ tvaṁ gacchasi)	किमर्थं त्वं गच्छसि अस्माभिः मे सम्मुखं प्रसारितं भविष्यति (kimarthaṁ tvaṁ gacchasi asmābhiḥ mē sammukhaṁ prasāritaṁ bhaviṣyati) किमर्थं त्वं गच्छसि इत्यत उच्यते कुरुष्विति (kimarthaṁ tvaṁ gacchasi ityata ucyatē kuruṣviti) किमर्थं त्वं गच्छसि तद्विदं विदे (kimarthaṁ tvaṁ gacchasi tadvidaṁ vidē)	किमर्थं त्वं गच्छसिन्नी गाजीवज़रहाडीज जिंदा कामर्न नामाजावन किमर्थं त्वं गच्छसियो का निराशा मामला चाहिए. इस त्वं गच्छसियो क किमर्थं त्वं गच्छसिनमः मरं त्वं से मृतमृत प्रदेशं प्रश्न□
मह्यं मिष्टान्नं रोचते (mahyaṁ miṣṭānnaṁ rōcatē)	मह्यं मिष्टान्नं रोचते ते ददामि ॥ १२॥ देवा अन्नं देहि मेऽजन्पिबन्ति। (mahyaṁ miṣṭānnaṁ rōcatē tē dadāmi .. 12.. dēvā annaṁ dēhi mē'janpibanti.) मह्यं मिष्टान्नं रोचते ते दिने दिने ॥ (mahyaṁ miṣṭānnaṁ rōcatē tē dinē dinē.) मह्यं मिष्टान्नं रोचते ते हि भक्त्या मया दत्तं वारिजाते सरोजे । (mahyaṁ miṣṭānnaṁ rōcatē tē hi bhaktyā mayā dattaṁ vārijātē sarōjē)	मह्यं मिष्टान्नं रोचते रहेंगे हैं, वो केंद्र में तुम निराश जि, मह्यं मिष्टान्नं रोचते तैयार भयरा ब्रेक्नां रोचते तैयार भयरा मह्यं मिष्टान्नं रोचते हुं ये क्यूँ मत है? मेरे आंदोलन बिल नह

Table 31: mParamanu 162M vs OPT 6.7B open end text generation for Sanskrit prompts.

Sanskrit Input Prompt	**Text Generation (top-3) from mParamanu 162M**	**LLaMa 2 7B**
अस्माकं देशः भारतवर्षः	अस्माकं देशः भारतवर्षः इति परिगणितः अस्माकं देशः भारतवर्षः इति ख्यातः अस्माकं देशः भारतवर्षः इति प्रसिद्धः	अस्माकं देशः भारतवर्षःअस्माकं देशः भारतवर्षः च, च निर्विषेषः श्रवणतिभिर्दिधिः ।रङ्ग अस्माकं देशः भारतवर्षः सहायकः वैयकितीय बीमानिक अर्थदान के आधार पर प्रशासनिक पात्रता पर व्याप <□> अस्माकं देशः भारतवर्षः / भारतचक्रवृक्षः। / महागृहः कालिश्छन्हितः । स्वप्नः स्वप्न'
वेदः चत्वारः सन्ति	वेदः चत्वारः सन्ति छन्दः ब्राह्मणम् तैत्तिरीयकम् कठोपनिषत् वैदिकच्छन्दसाम् मन्त्रम् आध्यपनम् अस्ति वेदः चत्वारः सन्ति प्रसिद्धः शिक्षाशास्त्रीयः शिक्षकायगः शास्त्रम् शास्त्रम् कणादः पाणिनीय शिक्ष काण्डे सूत्राणाम् उपरि भाष्यम् लिखितवान् वेदः चत्वारः सन्ति धर्मशास्त्राणि वेदाः वैदिककर्मकाण्डम् आन्वीक्षिकी दण्डनीतिप्रभृतीनि	वेदः चत्वारः सन्तिवेदः परः आरामवापेश्वर यूनरवापिताःचूत् वेदः चत्वारः सन्तिनःतान्यरकर्मः सर्वधिक्षित पर्वृति यत□ newline वेदः चत्वारः सन्ति चामिरः क्षमा दैविकौ । संकोषं च किं भारत निन्दा किं अतिनि । अपि मनुष्या
किमर्थं त्वं गच्छसि!	किमर्थं त्वं गच्छसि अस्माभिः मे सम्मुखं प्रसारितं भविष्यति किमर्थं त्वं गच्छसि इत्यत उच्यते कुरुष्विति किमर्थं त्वं गच्छसि तद्विदं विदे	किमर्थं त्वं गच्छसि रवित्वां हिंसे माया विच्छप्येकां गन्धसागरः ॥ ३ ॥ kīmartham tvam gacchasi ravitvā किमर्थं त्वं गच्छसि नित्यासङ्ग्रिमुत्थितम् । महेश्वरस्य महेश्वरति पश्यति परं तत्त्वम् किमर्थं त्वं गच्छसि यत् स्वप्रथमिति रतम्। 1-48-59,55-62 तस्मिन्नप्यहं मधुरं तपः
मह्यं मिष्टान्नं रोचते	मह्यं मिष्टान्नं रोचते ते ददामि ॥ १२॥ देवा अन्नं देहि मेऽजन्पिबन्ति। मह्यं मिष्टान्नं रोचते ते दिने दिने ॥ मह्यं मिष्टान्नं रोचते ते हि भक्त्या मया दत्तं वारिजाते सरोजे ।	मह्यं मिष्टान्नं रोचते मुविश्यते न्युइश्च यत्स उवाङ्ग्व्युं पश्यते तत्वः । तत्र नानिं श्वत मह्यं मिष्टान्नं रोचतेन्द्रोष्टतेन्द्रं राजन्यस्विंधादितिस्वीकृतिस्तत्रजगतिभारणि भावेशानि मृत् मह्यं मिष्टान्नं रोचते हिंता वा अनन्तकथ्या । पुत्रा कालाग्निः पृथिवी पदोंति मध्ये । स्वा यो

Table 32: mParamanu 162M vs LLaMa 2 7B open end text generation for Sanskrit prompts.

Sanskrit Input Prompt	Text Generation (top-3) from mParamanu 162M	GPTJ 6B
अस्माकं देशः भारतवर्षः	अस्माकं देशः भारतवर्षः इति परिगणितः अस्माकं देशः भारतवर्षः इति ख्यातः अस्माकं देशः भारतवर्षः इति प्रसिद्धः	अस्माकं देशः भारतवर्षः।। एतिहासपत्रः समाप्तःमैश्रीनिक् अस्माकं देशः भारतवर्षः। अथ मयि निबधमन्मा भारतयामि ते ।। □□□□□□□ □□□□□ □□ □□□ अस्माकं देशः भारतवर्षः रमाद्वी न गुणा ज्ञानं देशमस्तुलम् |व
वेदः चत्वारः सन्ति	वेदः चत्वारः सन्ति छन्दः ब्राह्मणम् तैत्तिरीयकम् कठोपनिषत् वैदिकच्छन्दसाम् मन्त्रम् आध्यपनम् अस्ति वेदः चत्वारः सन्ति प्रसिद्धः शिक्षाशास्त्रीयः शिक्षकायगः शास्त्रम् शास्त्रम् कणादः पाणिनीय शिक्ष काण्डे सूत्राणाम् उपरि भाष्यम् लिखितवान् वेदः चत्वारः सन्ति धर्मशास्त्राणि वेदाः वैदिककर्मकाण्डम् आन्वीक्षिकी दण्डनीतिप्रभृतीनि	वेदः चत्वारः सन्ति क्षिमेर्येणि भूतेर्वजाः स चतु भूत दुर्ग वेदः चत्वारः सन्तिभा सा कुण्डया चेति मम नो वो हृष्टी म् | योगिन□ वेदः चत्वारः सन्तिमाननाप्पतिः मुखमापरिणामहेमं पुनः सुमानः
किमर्थं त्वं गच्छसि	किमर्थं त्वं गच्छसि अस्माभिः मे सम्मुखं प्रसारितं भविष्यति किमर्थं त्वं गच्छसि इत्यत उच्यते कुरुष्विति किमर्थं त्वं गच्छसि तद्विदं विदे	किमर्थं त्वं गच्छसिं विस्मयम् । कुणात्वा नसिं मेदोः शायद कु□ किमर्थं त्वं गच्छसि) - अब हिंदु वै निम्रलाई बाँध्न हुने छ यस को किमर्थं त्वं गच्छसिद्धः (Thus I have declared and said, I will show thee what is my understanding of the meaning of these words.) कुप्तति स्मित्ता □
मह्यं मिष्टान्नं रोचते	मह्यं मिष्टान्नं रोचते ते ददामि ॥ १२॥ देवा अन्नं देहि मेऽजन्पिबन्ति। मह्यं मिष्टान्नं रोचते ते दिने दिने ॥ मह्यं मिष्टान्नं रोचते ते हि भक्त्या मया दत्तं वारिजाते सरोजे ।	मह्यं मिष्टान्नं रोचते शक्नोपने मितीन्तिस्तते वर्षते। मह्यं मिष्टान्नं रोचते O' My Soul! My Self is the Immortal. अष्टान्येन ततोऽस्त्यमात्म मह्यं मिष्टान्नं रोचते Innocent eyes are filled with tears, and the cheeks of the tender ones are wet with the tears of the heart, when a youth is told that his father is going to take away his life. But he forgets all those things, and only thinks of the joy of

Table 33: mParamanu 162M vs GPT-J 6B open end text generation for Sanskrit prompts.

Sanskrit Input Prompt	Text Generation (top-3) from mParamanu 162M	GPT-Neo 1.3B	GPT2-XL
अस्माकं देशः भारतवर्षः	अस्माकं देशः भारतवर्षः इति परिगणितः अस्माकं देशः भारतवर्षः इति ख्यातः अस्माकं देशः भारतवर्षः इति प्रसिद्धः	अस्माकं देशः भारतवर्षः। साक्षात्वायः। कारण वापरारित्याद्रद् अस्माकं देशः भारतवर्षः अस्माकं देशः केवल स्थापितः नका बहुत वि अस्माकं देशः भारतवर्षः सवै जातात्यायांन्ज़ी नावा तपासन्धुवः सु□	अस्माकं देशः भारतवर्षः स्थ्या कुण्यं में अस्माकं देशः भारतवर□ अस्माकं देशः भारतवर्षः वोडिया पायस्वहर्मिया ॥. One of the greatest benefits of being an e-book reader is that you don’t have to carry अस्माकं देशः भारतवर्षः भीषर्शिध्यता, योगमणिरणिः दो अमहानाति
वेदः चत्वारः सन्ति	वेदः चत्वारः सन्ति छन्दः ब्राह्मणम् तैत्तिरीयकम् कठोपनिषत् वैदिकच्छन्दसाम् मन्त्रम् आध्यपनम् अस्ति वेदः चत्वारः सन्ति प्रसिद्धः शिक्षाशास्त्रीयः शिक्षकायगः शास्त्रम् शास्त्रम् कणादः पाणिनीय शिक्ष काण्डे सूत्राणाम् उपरि भाष्यम् लिखितवान् वेदः चत्वारः सन्ति धर्मशास्त्राणि वेदाः वैदिककर्मकाण्डम् आन्वीक्षिकी दण्डनीतिप्रभृतीनि	वेदः चत्वारः सन्तिर्जनो गन्तु सेवेदः चत्वारः सन्तिर्जनो मि वेदः चत्वारः सन्तिसेंद्रिष्टीवरः स्तेमः स्त्रेमः स्त्रजा वेदः चत्वारः सन्तिवशावेदो चन्द्रादो श्री रं चवारास्रादो संस्	वेदः चत्वारः सन्तित्यार्गवेदंपर्यकारा नमर्थित्यार्गवेदं□ वेदः चत्वारः सन्तिर्थलस्पातिवर्तनाचान्नासद्धिरतंप्रविध□ वेदः चत्वारः सन्ति होगा, का सम्बेशर्च्छ्रजिंक्षिया अश्वेद च□
मह्यं मिष्टान्नं रोचते	मह्यं मिष्टान्नं रोचते ते ददामि ॥ १२॥ देवा अन्नं देहि मेऽजन्पिबन्ति। मह्यं मिष्टान्नं रोचते ते दिने दिने ॥ मह्यं मिष्टान्नं रोचते ते हि भक्त्या मया दत्तं वारिजाते सरोजे ।	मह्यं मिष्टान्नं रोचतेसारास्तान्नुसान्नश्च काव्याकसमान्ता□ मह्यं मिष्टान्नं रोचते। सँख्या भर्तरान्नराम्यनस्तेन। यसल मह्यं मिष्टान्नं रोचते त्रेभुचल्योगामस्रीयचिनःरुक्षणं	मह्यं मिष्टान्नं रोचते गाम्य वाकीच्छ हिंद्धिक्षिता महंति हो मह्यं मिष्टान्नं रोचतें ग्रमास्य भूखर्णव भुखवारोधिवर्णयम□ मह्यं मिष्टान्नं रोचतें समुनं कहतियों दूस्यया में रहीं में म
किमर्थं त्वं गच्छसि	किमर्थं त्वं गच्छसि अस्माभिः मे सम्मुखं प्रसारितं भविष्यति किमर्थं त्वं गच्छसि इत्यत उच्यते कुरुष्विति किमर्थं त्वं गच्छसि तद्विदं विदे	किमर्थं त्वं गच्छसि पर तर जहां तिथं त्वं गच्छिरण्यक्ष गतिवि□ किमर्थं त्वं गच्छसि तदतो अविजायिं वस्तु विदिमा आंतर्यतनोज किमर्थं त्वं गच्छसितो जातो किमर्थं त्वं प्रभ्यक्तिर्मित ग	किमर्थं त्वं गच्छसिथ्या, नोर्थ त्रह युजर प्रणं मुझ्यमापर कि□ किमर्थं त्वं गच्छसिर्णकार चेलार किलारी विष्णि प्रिशते चाहिय किमर्थं त्वं गच्छसिकार्व है। समझ्यत्रित्रान किम्मापस्त□

Table 34: mParamanu 162M vs GPT-Neo 1.3B vs GPT2-XL 1.5B open end text generation for Sanskrit prompts.

Sanskrit Input Prompt	mParamanu 162M	Bloom 560M	Bloom 1.1B	Bloom 3B
अस्माकं देशः भारतवर्षः	अस्माकं देशः भारतवर्षः इति परिगणितः अस्माकं देशः भारतवर्षः इति ख्यातः अस्माकं देशः भारतवर्षः इति प्रसिद्धः	अस्माकं देशः भारतवर्षः भारत (भारत सरकार)। भारत के संविधान में राज्य की राजधानी रांची को "राज्यसभा" में रखा गया है, जो रांची से लगभग 8 किलोमीटर की दूरी पर स्थित है। राज्यसभा के विस्तार में, रांची में एक "अस्माकं राष्ट्र" ((राज्यसभा) है। यह पूरे भारत की राजधानी में एक "राज्यसभा" का केंद्र है। राज्यसभा में भाग लेने के लिए सभी लोगों को अनिवार्य रूप से अपना नाम (अस्माकं देशः भारतवर्षः अस्माकं देशः भारतवर्षः चीनः ओमानः दक्षिण एशियाः ओज़ी एस्टेलीः ओमानः पाकिस्तानः प्रागैतिहासिकः प्रागैतिहासिकः अर्लाकं देशः प्रागैतिहासिकः सूरीः सूरीः सिलीगुड़ीः सिलीगुड़ीः सिलीगुड़ीः सिलीगुड़ीः सिलीगुड़ीः सिलीगुड़ीः सिलीगुड़ीः सिलीगुड़ीः सिलीगुड़ीः सिलीगुड़ीः	अस्माकं देशः भारतवर्षः अस्माकं देशः भारतवर्षः॥ ५ हे ! परमात्याय ! तत्रं तुस्य तत्रं॥ ७ हे ! परमात्माय ! तत्रं तुस्य तत्रं॥ ७ हे ! परमात्माय ! तत्रं तुस्य तत्रं॥ ७ हे ! परमात्माय ! तत्रं तुस्य तत्रं॥ ७ हे ! परमात्माय ! तत्रं तुस्य तत्रं॥ ७ नृत् अस्माकं देशः भारतवर्षः भीषर्शिध्यता, योगमणिरणिः दो अमहानाति	अस्माकं देशः भारतवर्षः । प्- प्र : देश। अस्मार : अस्मात्, अर्थात् इस प्रकार। प्- प्र : प्र : अस्माकं देशः भारतवर्षः ॥ौटिल्य ने इस देश का विवरण देने के लिए चार मुख्य लक्षणों के आधार पर अपनी रचना की है । अस्माकं देशः भारतवर्षः' आदि ने इसे स्वीकार किया। परंतु जब उसने सुना कि 'देवः लोकः' (देवलोक) का नाम प्रचलित हो चुका है, तब उसने इसे अस्वीकार कर दिया। इस प्रकार यह स्पष्ट है कि वैदिक काल से लेकर आधुनिक युग तक इस शब्द का प्रयोग बहुत ही प्रचलित है। संस्कृत और अवधी में इसकी अनेक शब्दावलियाँ प्रचलित हैं, जो कि वैदिक अर्थों में भिन्न हैं। पर उनमें मूल के समान अर्थ समाहित हैं। संस्कृत में', 'अस्माकं देशः भारतवर्षः । प्- प्र : देश। अस्मार : अस्मात्, अर्थात् इस प्रकार। प्- प्र : प्र : देश
वेदः चत्वारः सन्ति	वेदः चत्वारः सन्ति छन्दः ब्राह्मणम् तैत्तिरीयकम् कठोपनिषत् वैदिकच्छन्दसाम् मन्त्रम् आध्यपनम् अस्ति वेदः चत्वारः सन्ति प्रसिद्धः शिक्षाशास्त्रीयः शिक्षकायगः शास्त्रम् शास्त्रम् कणादः पाणिनीय शिक्ष काण्डे सूत्राणाम् उपरि भाष्यम् लिखितवान् वेदः चत्वारः सन्ति धर्मशास्त्राणि वेदाः वैदिककर्मकाण्डम् आन्वीक्षिकी दण्डनीतिप्रभृतीनि	वेदः चत्वारः सन्तिः - : - ! वेदः चत्वारः सन्ति समन् - नं द्वि ब , ग , - , - , - , नं ग , - , - , - , - , नं , ग - , - , - , - , नं , ग - , - , - , - , नं , ग - , - , - , - , नं ग - , - , - , - , नं , ग - , - , - , - , नं - वेदः चत्वारः सन्ति श्रष्टि प्रः त्र्याण् समि सर्वः देवानाः द्योति सर्वः सिः नः वं सः देवानाः द्योति सर्वः सिः नः वं सः देवानाः द्योति सर्वः सिः नः वं सः देवानाः द्योति सर्वः सिः नः वं सः देवानाः द्योति सर्वः सिः नः वं	वेदः चत्वारः सन्ति शल्लि । नः शांक्षते । वेदः चत्वारः सन्ति । नित्यानन्द- जो किसी भक्त की भक्ति करता है और भक्त के लिए कोई काम करता है उस भक्त का नाम भक्त ही है। इस बात को समझना चाहिए कि भक्त शब्द का अर्थ सिर्फ "जो" या "और" है। ये शब्द एक न्यास या साधन की बात है। भक्त की आत्मा के प्रति लगाव और भक्ति भावना ही भक्त का अर्थ है। यही भाव भक्त को परम पिता की कृपा से अपने भक्त के रूप', 'वेदः चत्वारः सन्ति ॥ ९ १ ॥ जो सत्य , परमात्मा को ही जानती है , वह भक्त है वेदः चत्वारः सन्ति नर्भि॥ ७८॥ नकुलः पर्माण्यश्च कत्रि॥ ७९॥ नकुलः नर्भे सत्वत्वः भृजः॥ ७७०॥ नकुलः नर्भे सर्वः चृथ्यः॥ ७ ७ १॥ नकुलः नर्भे पुरुषः नर्भे॥ ७ ७ २॥ नकुलः नर्भे शारीरः चक्ष्यः॥ ७ ७ ३	वेदः चत्वारः सन्ति प्रेयसी वेदः चत्वारः सन्ति . वेदः चत्वारः सन्ति प्रेयसी ।
महां मिष्टान्नं रोचते	महां मिष्टान्नं रोचते ते ददामि ॥ १२॥ देवा अन्नं देहि मेऽजन्पिबन्ति। महां मिष्टान्नं रोचते ते दिने दिने ॥ महां मिष्टान्नं रोचते ते हि भक्त्या मया दत्तं वारिजाते सरोजे ।	महां मिष्टान्नं रोचतेसारास्तान्नुसान्नश्च काव्याकसमान्ता महां मिष्टान्नं रोचते। सँख्या भर्तरान्नराम्यनस्तेन। यसल महां मिष्टान्नं रोचते त्रेभुचल्योगामसीयचिनः-रुक्षणं	महां मिष्टान्नं रोचते गाम्य वाकीच्छ हिंद्धिक्षिता महंति हो महां मिष्टान्नं रोचतें ग्रमास्य भूखर्णव भुखवारोधिवर्णयम महां मिष्टान्नं रोचतें समुनं कहतियों दूस्या में रहीं में म	अस्माकं देशः भारतवर्षः। साक्षात्वायः। कारण वापरारित्याद्रद् अस्माकं देशः भारतवर्षः अस्माकं देशः केवल स्थापितः नका बहुत वि अस्माकं देशः भारतवर्षः सवै जातात्यायांन्ज़ी नावा तपासन्धुवः सु
किमर्थं त्वं गच्छसि	किमर्थं त्वं गच्छसि अस्माभिः मे सम्मुखं प्रसारितं भविष्यति किमर्थं त्वं गच्छसि इत्यत उच्यते कुरुष्विति किमर्थं त्वं गच्छसि तद्विदं विदे	किमर्थं त्वं गच्छसि पर तर जहां तिथं त्वं गच्छिरण्यक्ष गतिवि किमर्थं त्वं गच्छसि तदतो अविजायिं वस्तु विदिमा आंतर्यतनोज किमर्थं त्वं गच्छसितो जातो िमर्थं त्वं प्रभ्यक्तिर्मित ग	किमर्थं त्वं गच्छसिथ्या, नोर्थ त्रह युजर प्रणं मुझ्यमापर कि किमर्थं त्वं गच्छसिर्णकार चेलार किलारी विष्णि प्रिशते चाहिय किमर्थं त्वं गच्छसिकार्व है। समझ्यत्रित्रान किम्मापस्त	अस्माकं देशः भारतवर्षः। साक्षात्वायः। कारण वापरारित्याद्रद् अस्माकं देशः भारतवर्षः अस्माकं देशः केवल स्थापितः नका बहुत वि अस्माकं देशः भारतवर्षः सवै जातात्यायांन्ज़ी नावा तपासन्धुवः सु

Table 35: mParamanu 162M vs Bloom (560M, 1.1B, 3B) open end text generation for Sanskrit prompts.

Hindi Input Prompt	Text Generation (top-3) from Paramanu-Hindi 162M	GPT-Neo 1.3B	GPT2-XL
सचिन तेंदुलकर (Sachin Tendulkar)	सचिन तेंदुलकर, गांगुली, द्रविड़, लक्ष्मण और सहवाग के अलावा विराट कोहली और आर अश्विन के 18-18 छक्कों के रिकॉर्ड की भी बीसीसीआई ने खोज शुरू कर दी है। (sacina tēṁdulakara, gāṁgulī, draviṛa, lakṣmaṇa aura sahavāga kē alāvā virāṭa kōhalī aura āra aśvina kē 18-18 chakkōṁ kē rikôrḍa kī bhī bīsīsīāī nē khōja śurū kara dī hai.) सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है। (sacina tēṁdulakara kā ākhirī va 200 vāṁ ṭēsṭa kāphī ahama hai aura isē ēka mahāna khilāṛī kō yādagāra vidāī dēnē kē lihāja sē bēhada ahama kahā jā sakatā hai.) सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है। (sacina tēṁdulakara kā ākhirī va 200 vāṁ ṭēsṭa kāphī ahama hai aura isē ēka mahāna khilāṛī kō yādagāra vidāī dēnē kē lihāja sē bēhada ahama kahā jā sakatā hai.)	सचिन तेंदुलकर एक भावने पर चारने से बाहर है यहाँ के रंग से यहाँ स्थ सचिन तेंदुलकर सुवना की जगह की महाराज तिरस्मादित्या विशेष राव्य सचिन तेंदुलकर्ताओं के रूप में एक वैश्विक करना है, इसलिए हमें य□	सचिन तेंदुलकर्थविगति में सम्हंते मवाचन भववामन है दियाल का गय□ सचिन तेंदुलकरोप्रवेद्धितं यही के जम भविश्वं के बाद जिंग के □ सचिन तेंदुलकर्मों दोनीदियंत्र तुशांके यहकल आना दी भीषर्शित
शाहरुख़ ख़ान (Shah Rukh Khan)	शाहरुख़ ख़ान की फ़िल्म डॉन को आज मुंबई और आसपास के इलाकों में बड़े पैमाने पर देखा जा रहा है। (śāharuk□ha k□hāna kī filma ḍôna kō āja muṁbaī aura āsapāsa kē ilākōṁ mēṁ baṛē paimānē para dēkhā jā rahā hai.) शाहरुख़ ख़ान, रीतेश देशमुख, अनुष्का शर्मा, रणवीर शौरी, लारा दत्ता, नाना पाटेकर, आशीष चौधरी, नवजोत सिद्धू, शक्ति कपूर, गोविंदा, चंकी पांडे के अभिनय वाली इस फिल्म ने शुक्रवार को 10.11 करोड़ की कमाई की। (śāharuk□ha k□hāna, rītēśa dēśamukha, anuṣkā śarmā, raṇavīra śaurī, lārā dattā, nānā pāṭēkara, āśīṣa caudharī, navajōta siddhū, śakti kapūra, gōviṁdā, caṁkī pāṁḍē kē abhinaya vālī isa philma nē śukravāra kō 10.11 karōṛa kī kamāī kī.) शाहरुख़ ख़ान के प्रशंसक बनें या शाहरुख़ ख़ान के विरोधी , शाहरुख़ ख़ान को लेकर बहुत कुछ लिखा जा सकता है (śāharuk□ha k□hāna kē praśaṁsaka banēṁ yā śāharuk□ha k□hāna kē virōdhī , śāharuk□ha k□hāna kō lēkara bahuta kucha likhā jā sakatā hai)	शाहरुख़ ख़ानी चापलबुती रिलाई वापरते मेरा बरेशति के शिक्षता सीज शाहरुख़ ख़ाने ही रहते हैं। अब तुम्हें पता लगाने के लिए मुख्य का व' शाहरुख़ ख़ाने पर भीतर पिछले सप्ताह के रूप में जीवन चालू करने के	शाहरुख़ ख़ान खादने प्रहारुख़ भारत करणे है तो तोले के उस के रोजी भा शाहरुख़ ख़ाने लिए अयोगाने वाली वरीज़ा दोववे मन्दी लोगें पहले क शाहरुख़ ख़ान कोई यही मनीजर के लोगा आने है?करी नहीं की साथ रहता कर
महात्मा गांधी (Mahatma Gandhi)	महात्मा गांधी और उनके समकालीन नेताओं का मुख्य ध्यान समाज के अंतिम व्यक्ति को महत्व देने पर था महात्मा गांधी राष्ट्रीय ग्रामीण रोजगार गारंटी योजना के तहत संविदा के कर्मचारियों की हड़ताल का असर इन्दौर में भी दिखा। महात्मा गांधी भी अपनी जिंदगी में मर्यादा का पालन करते थे और अगर वे हिंसा का सहारा ले रहे हैं तो वे भी विचारधारा के शिकार हुए हैं।	महात्मा गांधी महात्मा गांधी के भावनामधील आनंद भर आवश्यकता के राजीव्य गिनी गांध' महात्मा गांधी के भावनामधील आनंद भर आवश्यकता के राजीव्य गिनी गांध महात्मा गांधीशोऽते गार्जी आवश्यक आहे. तर तथा व्यवरण नामावरणो हन	महात्मा गांधी भी साथा पूरवस बराहमें शोहिलेकंपरत्या हैं Name:Phone No: महात्मा गांधी की से करेंड़ नीते है मुस्ती और दिया जानके से हमें आ□ महात्मा गांधी का में समग्रवोरवो देखारा से पर पर रहाते है, जिस्कार □
लता मंगेशकर (Lata Mangeshkar)	लता मंगेशकर नूरजहां और शमशाद बेगम के साथ भी गा चुकी थीं। (latā maṁgēśakara nūrajahāṁ aura śamaśāda bēgama kē sātha bhī gā cukī thīṁ.) लता मंगेशकर का जन्म २८ सितम्बर १९२९ को वर्तमान पाकिस्तान में हुआ था। (latā maṁgēśakara kā janma 28 sitambara 1929 kō vartamāna pākistāna mēṁ huā thā.) लता मंगेशकर ने रील लाइफ से रियल लाइफ तक का सफर तय किया है और रियल लाइफ के उनके करीबियों ने उन्हें रील लाइफ में भी न सिर्फ रियल लाइफ बल्कि लव लाइफ में भी जीना सिखा दिया है। (latā maṁgēśakara nē rīla lāipha sē riyala lāipha taka kā saphara taya kiyā hai aura riyala lāipha kē unakē karībiyōṁ nē unhēṁ rīla lāipha mēṁ bhī na sirpha riyala lāipha balki lava lāipha mēṁ bhī jīnā sikhā diyā hai.)	लता मंगेशकर है तब मेरा नाम गुल्मने में अक्षम कर दिया है यह वास्तव लता मंगेशकर करने की जरूरत हो जाए तेल्स के अंदर बाहर अमेज में रखा ज लता मंगेशकर अपेक्षा में हों या उसका दूसरा एक सारा हजारों से स्पष्ट	लता मंगेशकर प्रायना को बहुतो गयाते समेश्वे दिलीनने गया विचने क लता मंगेशकर स्वास रहित आरवोईरहे मन्दुनीते के नहीं सब मंगेशकर लता मंगेशकर प्रयवेस प्रयावसने जानकारुष्यियतेत्के देश सिवास

Table 36: Paramanu-Hindi 162M vs GPT-Neo 1.3B vs GPT2-XL 1.5B open end text generation for Hindi prompts.

Hindi Input Prompt	**Text Generation (top-3) from Paramanu-Hindi 162M**	**OPT 6.7B**
सचिन तेंदुलकर (Sachin Tendulkar)	सचिन तेंदुलकर, गांगुली, द्रविड़, लक्ष्मण और सहवाग के अलावा विराट कोहली और आर अश्विन के 18-18 छक्कों के रिकॉर्ड की भी बीसीसीआई ने खोज शुरू कर दी है। सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है। सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है।	सचिन तेंदुलकर क्षेत्र रिपोर्ट I am a writer and a thinker. I write on contemporary political, social, cultural, economic, psychological, and philosophical issues. I write on the subjects of politics, philosophy, psychology, and sociology. I am an amateur सचिन तेंदुलकर आयोजित कम संयुक्त करेगा।लेकिन जब संयुक्त कम सचिन तेंदुलकर एक दो वह पास करते रहे हैं आज मगर अब पास करते हैं सब
शाहरुख़ ख़ान (Shah Rukh Khan)	शाहरुख़ ख़ान की फ़िल्म डॉन को आज मुंबई और आसपास के इलाकों में बड़े पैमाने पर देखा जा रहा है। शाहरुख़ ख़ान, रीतेश देशमुख, अनुष्का शर्मा, रणवीर शौरी, लारा दत्ता, नाना पाटेकर, आशीष चौधरी, नवजोत सिद्धू, शक्ति कपूर, गोविंदा, चंकी पांडे के अभिनय वाली इस फिल्म ने शुक्रवार को 10.11 करोड़ की कमाई की। शाहरुख़ ख़ान के प्रशंसक बनें या शाहरुख़ ख़ान के विरोधी , शाहरुख़ ख़ान को लेकर बहुत कुछ लिखा जा सकता है	शाहरुख़ ख़ानी चापलबुती रिलाई वापरते मेरा बरेशति के शिक्षता सीज□ शाहरुख़ ख़ाने ही रहते हैं। अब तुम्हें पता लगाने के लिए मुख्य का व' शाहरुख़ ख़ाने पर भीतर पिछले सप्ताह के रूप में जीवन चालू करने के □
महात्मा गांधी (Mahatma Gandhi)	महात्मा गांधी और उनके समकालीन नेताओं का मुख्य ध्यान समाज के अंतिम व्यक्ति को महत्व देने पर था महात्मा गांधी राष्ट्रीय ग्रामीण रोजगार गारंटी योजना के तहत संविदा के कर्मचारियों की हड़ताल का असर इन्दौर में भी दिखा। महात्मा गांधी भी अपनी जिंदगी में मर्यादा का पालन करते थे और अगर वे हिंसा का सहारा ले रहे हैं तो वे भी विचारधारा के शिकार हुए हैं।	महात्मा गांधी राम की अनुराष्ट्रित बल नहीं, आनंद को खोज के आरोपि□ महात्मा गांधी ने मानवरी अपना प्रवेश के नेतृत्व को बढ़ाने के लिए □ महात्मा गांधी में मैंने नागलिंग देखने को सेना की संपर्क के साथ आ□
लता मंगेशकर (Lata Mangeshkar)	लता मंगेशकर नूरजहां और शमशाद बेगम के साथ भी गा चुकी थीं। लता मंगेशकर का जन्म २८ सितम्बर १९२९ को वर्तमान पाकिस्तान में हुआ था। लता मंगेशकर ने रील लाइफ से रियल लाइफ तक का सफर तय किया है और रियल लाइफ के उनके करीबियों ने उन्हें रील लाइफ में भी न सिर्फ रियल लाइफ बल्कि लव लाइफ में भी जीना सिखा दिया है।	लता मंगेशकरती अमेरा यांची बाला सोमवार यांची जनता का सही दिवस शब्द' लता मंगेशकर पीसद में एक दिनों प्राप्त पहुंचे। इस निधन से पहुं लता मंगेशकर आज अंदर बालिक बने विधायक पर कब्जे को देखा हूँ। इसलि

Table 37: Paramanu-Hindi 162M vs OPT 6.7B open end text generation for Hindi prompts.

Hindi Input Prompt	Text Generation (top-3) from Paramanu-Hindi 162M	LLaMa-2 7B
सचिन तेंदुलकर (Sachin Tendulkar)	सचिन तेंदुलकर, गांगुली, द्रविड़, लक्ष्मण और सहवाग के अलावा विराट कोहली और आर अश्विन के 18-18 छक्कों के रिकॉर्ड की भी बीसीसीआई ने खोज शुरू कर दी है। सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है। सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है।	सचिन तेंदुलकर सचिन तेंदुलकर के हकीकतें जानिए सचिन तेंदुलकर के हकीकतें जानिए और करें सही घोषणा. प्रमुख मुद्दों पर राजे पाली पहले इशारे दिए। बॉलीवुड में रहने वाले सितारों में कोई फेहरा नहीं बल्कि बेहद सुखदायी है। फिर भी सब मानेंगे सितारों को उम्मीद सचिन तेंदुलकर राउंड में मार रहे कुछ विदेशियों के प्लेवर ने दिल झका लिया, बीस फेमिना की मारी रही स्पिक्स स्कुल पाठ्यालय के एडमिनिस्ट्रेट फांदीन अहमद जी की जांच पर दिल्ली में कई मुस्लिम महिलाओं को हमला करने की आकलाम आयीं । बहुत जल्दी बाद
शाहरुख़ ख़ान (Shah Rukh Khan)	शाहरुख़ ख़ान की फ़िल्म डॉन को आज मुंबई और आसपास के इलाकों में बड़े पैमाने पर देखा जा रहा है। शाहरुख़ ख़ान, रीतेश देशमुख, अनुष्का शर्मा, रणवीर शौरी, लारा दत्ता, नाना पाटेकर, आशीष चौधरी, नवजोत सिद्धू, शक्ति कपूर, गोविंदा, चंकी पांडे के अभिनय वाली इस फिल्म ने शुक्रवार को 10.11 करोड़ की कमाई की। शाहरुख़ ख़ान के प्रशंसक बनें या शाहरुख़ ख़ान के विरोधी, शाहरुख़ ख़ान को लेकर बहुत कुछ लिखा जा सकता है	शाहरुख़ ख़ान और अमिताब बच्चन ने रोहित शर्मा के निकाट रहे सिलवाई शादी में अनुभव की दुनिया देखा Dear reader, its been a little over a month since Salman Khans hit-and-run case was closed. While a lot of speculations had been doing the rounds about Salmans alleged affair with Aishwarya Rai, the superstar has repeatedly denied it and said that he is a happily married man. Well, Salmans wife, Arpita Khan, has been very supportive of her husband. In fact, its been widely reported that Arpita and Salmans sister, Alvira, have been quite worried about the allegations. They have been concerned about the reports about the superstars affairs with women like Katrina Kaif, A", '<s> शाहरुख़ ख़ान की बॉलीवुड करियर के संकटों से रिकॉर्ड क्रिकेटर लेक कहा कि यह सौभाग्य था कि मेरे पिता महादेव कपूर नहीं थे और हम नहीं मिलीं। रिकॉर्ड क्रिकेटर लेक खिलाड़ी पर बढ़ते हुए क्रिकेटर के एक बार फिर नाम बाज़ार में आ गये। ये बात सबके स शाहरुख़ ख़ान की बॉलीवुड करियर के संकटों से रिकॉर्ड क्रिकेटर लेक कहा कि यह सौभाग्य था कि मेरे पिता महादेव कपूर नहीं थे और हम नहीं मिलीं। रिकॉर्ड क्रिकेटर लेक खिलाड़ी पर बढ़ते हुए क्रिकेटर के एक बार फिर नाम बाज़ार में आ गये। ये बात सबके स शाहरुख़ ख़ान ने संगम अलफ़ाजी की निशानी के खिलाफ घेर दिया हाथ 2023-03-25T15:36:04+05:30 Shah Rukh Khans Mother Shares Video Of Her Husband "Taking Sip Of Water After Taking His Medicine" Shah Rukh Khan's mother, Lateef Fatima Khan, took to her Instagram handle to share a video of her husband "taking a sip of water after taking his medicine".Recently, Sharukh Khan shared a heartfelt post for his father. He also shared a video of him in which he is seen speaking on his fathers health. In the clip, SRK can be seen sitting in a chair and holding his fathers hand. The video went viral on social media as many people were moved by the sight of the King Khan holding his fathers hand. SRK had
महात्मा गांधी (Mahatma Gandhi)	महात्मा गांधी और उनके समकालीन नेताओं का मुख्य ध्यान समाज के अंतिम व्यक्ति को महत्व देने पर था महात्मा गांधी राष्ट्रीय ग्रामीण रोजगार गारंटी योजना के तहत संविदा के कर्मचारियों की हड़ताल का असर इन्दौर में भी दिखा। महात्मा गांधी भी अपनी जिंदगी में मर्यादा का पालन करते थे और अगर वे हिंसा का सहारा ले रहे हैं तो वे भी विचारधारा के शिकार हुए हैं।	महात्मा गांधी अंको अपलाईल होली पाउडाको सुवासियालाई लिनक मुख्य राष्ट्रपतिमा पत्र मिला Kathmandu, July 25 (RSS): Rashtriya Samajbadi Party Nepal (RSPN) leader Madhav Kumar Nepal has written a letter to President Bidya Devi Bhandari seeking the appointment of former prime ministers Sher Bahadur Deuba and Pushpa Kamal Dahal as the prime minister, RSPNs central office said. In the letter dated 24 July, Nepal requested the president to make Nepal Communist Party (NCP) chairperson and former prime minister Deuba, RSPN general secretary and former prime minister Dahal and NCP leader and former prime minister Jhala Nath Khanal prime minister through a joint consensus. The महात्मा गांधी की परिचितता बढ़ाने के लिए सुधार प्रक्रिया का आह्वान अभियान की शुरुआत करें श्रीमान महात्मा गांधी की मान्यता हिन्दुस्तान की सर्वोच्चता प्रेम की सीमाओं से बाहर से अच्छी तरह से अतिरिक्त होना चाहिए और हिन्दुस्तान पर एक दूसरी ध्यान महात्मा गांधी का साहित्यिक पत्रिका भारतीय लोकसभा महासभा में कौन वाले बड़े राजनितिक प्रधान थे? कबीर सिंह की बेटी सत्य सिंह को कब समेत देश की महान तरीक़ा प्रेरित होती थीं? उसकी परिवार व घर के जांच से कहीं नहीं पाई जाती थी कि जातिवाद का गुलाम अब्दुल
लता मंगेशकर (Lata Mangeshkar)	लता मंगेशकर नूरजहां और शमशाद बेगम के साथ भी गा चुकी थीं। लता मंगेशकर का जन्म २८ सितम्बर १९२९ को वर्तमान पाकिस्तान में हुआ था। लता मंगेशकर ने रील लाइफ से रियल लाइफ तक का सफर तय किया है और रियल लाइफ के उनके करीबियों ने उन्हें रील लाइफ में भी न सिर्फ रियल लाइफ बल्कि लव लाइफ में भी जीना सिखा दिया है।	लता मंगेशकर पर लता मंगेशकर बोले हर पहले उच्च शिक्षा के बारे में अपनी दुर्घटना से सम्बंधित बात-चीत लता मंगेशकर की बेटी पियाली ने बताया कहां की उम्र से शादी करना चाहते हैं प्याली एक अदृश्य विवाहिक नाम बता रही है। वह अपने किस्सें अकेला खासी एवं अपोकेतिक अर्थात्मक होने पर वापस अपनी प्याली पर कहती हैं। पात्रता से प्याली उम्र छात्र-दारियों से म

Table 38: Paramanu-Hindi 162M vs LLaMa-2 7B open end text generation for Hindi prompts.

Hindi Input Prompt	Text Generation (top-3) from Paramanu-Hindi 162M	GPT-J 6B
सचिन तेंदुलकर (Sachin Tendulkar)	सचिन तेंदुलकर, गांगुली, द्रविड़, लक्ष्मण और सहवाग के अलावा विराट कोहली और आर अश्विन के 18-18 छक्कों के रिकॉर्ड की भी बीसीसीआई ने खोज शुरू कर दी है। सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है। सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है।	सचिन तेंदुलकर वरचेता जाते. Kannada: ಹೆಚ್ಚುವರಿ ತಿದ್ದನೆಗಳು ವಾದ ಮತ್ತು ಸತ್ಯ ಎಂದು ಕರೆದಾಗುತ್ತಾರೆ. Serbian: I am a writer and a thinker. I write on contemporary political, social, cultural, economic, psychological, and philosophical issues. I write on the subjects of politics, philosophy, psychology, and sociology. I am an amateur सचिन तेंदुलकर शेखपद तकियोवा □□□□□□□ डॉवरका शनिवार, 10 जनवरी, 2010 को दंपत दूल्हें दो बार दी सिंहसिंग देखभाल और कीटनाशकों में तरीका समाप्त। अगले देर में दो बार देखभाल और बचाव सचिन तेंदुलकर हैन्च होकिन्ने दुष्पतेंदुलकर होकिन्ने दुष्पतेंदुलकर होकिन्ने दुष्पतेंदुलकर होकिन्ने दुष्पतेंदुलकर होकिन्ने दुष्पतेंदुलकर होकिन्
शाहरुख़ ख़ान (Shah Rukh Khan)	शाहरुख़ ख़ान की फ़िल्म डॉन को आज मुंबई और आसपास के इलाकों में बड़े पैमाने पर देखा जा रहा है। शाहरुख़ ख़ान, रीतेश देशमुख, अनुष्का शर्मा, रणवीर शौरी, लारा दत्ता, नाना पाटेकर, आशीष चौधरी, नवजोत सिद्धू, शक्ति कपूर, गोविंदा, चंकी पांडे के अभिनय वाली इस फिल्म ने शुक्रवार को 10.11 करोड़ की कमाई की। शाहरुख़ ख़ान के प्रशंसक बनें या शाहरुख़ ख़ान के विरोधी , शाहरुख़ ख़ान को लेकर बहुत कुछ लिखा जा सकता है	शाहरुख़ ख़ानेकु काके गहु होइनको परम्परामाले जस्तो छैन शेल्गा काकेकु शाहरुख़-रूपमा समाजको अरूलाई हाम्रो रोजिसम्म यति नयाँ शहरको बचाउनु परंतु यसको जनबीतको जाति□ शाहरुख़ ख़ानेर तिमिरन कहुन्छ यस्तोमा खुटेको लागि काम गर्नु कर्ने हुँदै Estonian: Nagu siin on hästi, saate teadada, kuidas mina kirjutasin sinise koos tundumata, kuidas kirjutasin ülejäänud kõikide aegade tundumata. Nagu siin on hästi, saate teadada, kuidas mina kirjutasin sinise koos tundumata, kuidas kirjutasin ülejäänud kõikide aegade tundumata. शाहरुख़ ख़ानाहरु देखाको थिए तीन बाँयालहरुको विवाहमा पनि पैसा लगाएका थिए texthindi Turkish: Japanese: インドの女性は若いときにお金を使って食べました 三鳥のような子供はお金を取得することができます 結婚指輪は小さな社会で支払われました インドでは□
महात्मा गांधी (Mahatma Gandhi)	महात्मा गांधी और उनके समकालीन नेताओं का मुख्य ध्यान समाज के अंतिम व्यक्ति को महत्व देने पर था महात्मा गांधी राष्ट्रीय ग्रामीण रोजगार गारंटी योजना के तहत संविदा के कर्मचारियों की हड़ताल का असर इन्दौर में भी दिखा। महात्मा गांधी भी अपनी जिंदगी में मर्यादा का पालन करते थे और अगर वे हिंसा का सहारा ले रहे हैं तो वे भी विचारधारा के शिकार हुए हैं।	महात्मा गांधी प्राप्तमुपाते, सिद्ध समृद्ध भारतीयेत्र ब्रह्माच्छिन्न महात्मा पुरुष स्वतः हैतो। वहाँ तो केहि भी अज्ञा आयत आहे जेसे शते व भैंके, पुन्हा शेष भैंके महात्मा गांधी में मृत के वर्ष में पहली बार कितना परेशानियां हुई। उन्होंने देखा है कि एक युवती का पैर अपनी देखरी पीठ में लहसुन भूमिका नहीं करती है। बाबा ने पीड़ित सोने महात्मा गांधी की रोक करने की कोशिश करती हैं। अपनी पसंद व्यक्त करती हैं। अपनी देहिया के शरीर में इसे बदलना नहीं चाहती हैं। सेना को असुविधा दी जिसे सिर्फ अधिक न
लता मंगेशकर (Lata Mangeshkar)	लता मंगेशकर नूरजहां और शमशाद बेगम के साथ भी गा चुकी थीं। लता मंगेशकर का जन्म २८ सितम्बर १९२९ को वर्तमान पाकिस्तान में हुआ था। लता मंगेशकर ने रील लाइफ से रियल लाइफ तक का सफर तय किया है और रियल लाइफ के उनके करीबियों ने उन्हें रील लाइफ में भी न सिर्फ रियल लाइफ बल्कि लव लाइफ में भी जीना सिखा दिया है।	लता मंगेशकर को मालते हुए आश्वस्तेज तरंगी का तेजतर पैगंबन बन गया है। उन्होंने पिछले 5 कोबीर बना मुख्य दुल्हन के अनुभाग में समाप्त किया है। मैनेजिक मोल्डन के एक कह लता मंगेशकर्ता द्वारा भुगतान के पूरी तरह से कनेक्शन को समाप्त कर दिया। तो अब मैं आपको बताना चाहूंगा कि कौन सा व्यक्ति ने किसी चीज़ का भुगतान किया है? आपको एक दूसरे की भ लता मंगेशकर्षण का भेंकं को हमेशा तो जनाती होता हैं। लेकिन वे काफी खूबसूरते हैं। Portuguese: Eu tenho minha própria música de fundo. Eu tenho um grande amigo em que o meu nome significa muito para ele, e ele fala em mim no último episódio de uma maneira realmente agradável, e é muito bom se você acompanhar isso, e ele é Muito boa como o meu amigo, ele é meu amigo, como o meu amigo, e ele é tão

Table 39: Paramanu-Hindi 162M vs GPT-J 6B open end text generation for Hindi prompts.

Hindi Input Prompt	Text Generation (top-3) from Paramanu-Hindi 367.5M	Bloom 560M	Bloom 1.1B	Bloom 3B	Bloom 7.1B
सचिन तेंदुलकर	सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है। सचिन तेंदुलकर का आखिरी व 200 वां टेस्ट काफी अहम है और इसे एक महान खिलाड़ी को यादगार विदाई देने के लिहाज से बेहद अहम कहा जा सकता है। सचिन तेंदुलकर, गांगुली, द्रविड़, लक्ष्मण और सहवाग के अलावा विराट कोहली और आर अश्विन के 18-18 छक्कों के रिकॉर्ड की भी बीसीसीआई ने खोज शुरू कर दी है।	सचिन तेंदुलकर ने कहा है कि उन्हें ये बात पसंद नहीं है कि भारतीय टीम में किसी खिलाड़ी की जगह दूसरी पारी में खेलने का मौका न मिल रहा हो । सचिन तेंदुलकर ने कहा है कि उन्हें ये बात पसंद नहीं है कि भारतीय टीम में किसी खिलाड़ी की जगह दूसरी पारी में खेलने का मौका न मिल रहा हो।	सचिन तेंदुलकर से मुलाकात करेंगे और अपने नए कप्तानों को सम्मानित करेंगे। सचिन तेंदुलकर और भारतीय कप्तान महेंद्र सिंह धोनी के बीच चल रही विवाद के बीच अब ऑस्ट्रेलिया के टेस्ट कप्तान माइकल राइट ने भारतीय कप्तान को 'मीडिया के जिम्मेदार' बताया है।	सचिन तेंदुलकर के खिलाफ भारतीय टीम के पूर्व कप्तान महेंद्र सिंह धोनी ने कहा है कि उन्होंने टीम को जीत दिलाने के लिए एक टीम की जरूरत नहीं महसूस की सचिन तेंदुलकर और महेंद्र सिंह धोनी के लिए इस बात का मायने कई बार नहीं होता कि भारत के लिए खेलने वाले दो महान बल्लेबाज अब अपने करियर की समाप्ति पर अपने करियर को थाम लें	सचिन तेंदुलकर का मानना है कि सचिन तेंदुलकर को क्रिकेट का असली भगवान नहीं माना जा सकता क्योंकि उस समय सचिन की उम्र महज़ 19 वर्ष थी लेकिन वह खेल की दुनिया में छा गए और अब वह भारत में सबसे लोकप्रिय क्रिकेट खिलाड़ी है । सचिन तेंदुलकर के साथ है, जिन्होंने यह पुरस्कार जीता था। शारदा सिन्हा (सरस), पूर्व क्रिकेटर और टीवी पत्रकार शारदा सिन्हा (सरस), पूर्व क्रिकेटर और टीवी पत्रकार के रूप में जाना जाता है सचिन तेंदुलकर और वीरेंद्र सहवाग भी शामिल थे।
शाहरुख़ ख़ान	शाहरुख़ ख़ान, रीतेश देशमुख, अनुष्का शर्मा, रणवीर शौरी, लारा दत्ता, नाना पाटेकर, आशीष चौधरी, नवजोत सिद्धू, शक्ति कपूर, गोविंदा, चंकी पांडे के अभिनय वाली इस फिल्म ने शुक्रवार को 10.11 करोड़ की कमाई की। शाहरुख़ ख़ान की फ़िल्म डॉन को आज मुंबई और आसपास के इलाकों में बड़े पैमाने पर देखा जा रहा है। शाहरुख़ ख़ान ने थैंक्यू सॉरीज़ बोलते हुए कहा, मैने अपने सभी करीबी दोस्तों और परिवार वालों से कहा है कि ये समारोह बहुत ज़रूरी था।	शाहरुख़ ख़ान के साथ है, जिन्होंने एक फिल्म का निर्देशन किया है। शाहरुख़ ख़ान के साथ फिल्म के पहले दृश्य में अमिताभ बच्चन का हाथ और कान पीछे रखकर एक शर्ट से पकड़े जाने वाला है. शाहरुख़ ख़ान का शर्ट और पैंटी को देखकर अंदाजा लगाया जा सकता है कि वे बहुत देर से पैंट और शर्ट पहन रही थीं।	शाहरुख़ ख़ान को श्रद्धांजलि दी. शाहरुख़ ख़ान की बेटी सुहाना खान अपने पहले बर्थडे सेलिब्रेट कर रहीं हैं. सुहाना ने ट्विटर पर इस बात का ऐलान किया है कि वे अपनी पहली बर्थडे पार्टी 22 दिसंबर को दिल्ली के एक कैफे में कर रही हैं शाहरुख़ ख़ान एक बार फिर से अपनी नई फ़िल्म धूम 3 के लिए अपने परिवार संग इस शहर में आए हैं	शाहरुख़ ख़ान के अनुसार, उन्हें कभी नहीं लगा कि उनकी फ़िल्म को कोई गंभीरता से ले रहा है. लेकिन फ़िल्म से पहले उन्हें लगता था कि लोग उन्हें पसंद करेंगे. लेकिन असलियत में दर्शकों का रवैया बिल्कुल अलग था. शाहरुख़ ख़ान की	शाहरुख़ ख़ान के प्रशंसक बनें या शाहरुख़ ख़ान के विरोधी, शाहरुख़ ख़ान को लेकर बहुत कुछ लिखा जा सकता है
	लता मंगेशकर नूरजहां और शमशाद बेगम के साथ भी गा चुकी थीं। लता मंगेशकर का जन्म २८ सितम्बर १९२९ को वर्तमान पाकिस्तान में हुआ था। लता मंगेशकर ने रील लाइफ से रियल लाइफ तक का सफर तय किया है और रियल लाइफ के उनके करीबियों ने उन्हें रील लाइफ में भी न सिर्फ रियल लाइफ बल्कि लव लाइफ में भी जीना सिखा दिया है।	लता मंगेशकर। लता मंगेशकर ने लिखा है कि वह खुद की शादी को लेकर बिल्कुल भी शर्मशार नहीं है और वह अभी भी अपने लिए एक सुंदर और फिट मां बनना चाहती है। लता मंगेशकर , दीपक केसरिया , मनोज कुमार , शुभम शर्मा , अभिषेक शर्मा आदि नेगी	लता मंगेशकर , हरिओम चौहान , अमित शर्मा , विकास जैन , धर्मबीर जैन , सुनील शर्मा , बलबीर त्यागी , सतीश कुमार , राजेंद्र कुमार आदि मौजूद रहे । लता मंगेशकर की हत्या की जांच को शुरू किया गया है। बता दें कि ये मामला 17 साल पुराना है। इस मामले में पहले से दो आरोपी पुलिस की गिरफ्त में थे लेकिन पिछले दो महीने में यह मामला काफी बढ़ गया था। इससे पहले पूर्व मंत्री गायत्री प्रसाद प्रजापति के बेटे गायत्री प्रजापति और भाजपा विधायक संगीत सोम के बीच हुई झड़प के बाद अब गायत्री प्रजापति की हत्या की आशंका जताई जा रही है। आपको बता दें कि गायत्री प्रजापति की हत्या लता मंगेशकर द्वारा लिखा गया।	लता मंगेशकर ने कहा कि मुझे बहुत अच्छा लग रहा है कि इतने वर्षों के बाद मेरी आवाज में वह शायरी आ गई लता मंगेशकर , दिलीप ताहिल , विजय जाधव , मंगेश कदम , शशिकांत कदम , यशवंत शिंदबराव , प्रशांत ताहिल , विजय जाधव , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंद', 'लता मंगेशकर , मंगेश कदम , सुरेश पंडित , सचिन पिळगावकर , सुरेश वाडकर , लक्ष्मीकांत प्यारेलाल , अनिल विश्वास , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल लता मंगेशकर , मंगेश कदम , सुरेश पंडित , सचिन पिळगावकर , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल	लता मंगेशकर ने कहा कि मुझे बहुत अच्छा लग रहा है कि इतने वर्षों के बाद मेरी आवाज में वह शायरी आ गई लता मंगेशकर , दिलीप ताहिल , विजय जाधव , मंगेश कदम , शशिकांत कदम, यशवंत शिंदबराव, प्रशांत ताहिल, विजय जाधव, विजय शिंदबराव, यशवंत शिंदबराव, प्रशांत ताहिल , विजय शिंदबराव, यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव, यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव, यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव, यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव, यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव , यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंदबराव, यशवंत शिंदबराव , प्रशांत ताहिल , विजय शिंद', 'लता मंगेशकर , मंगेश कदम , सुरेश पंडित , सचिन पिळगावकर , सुरेश वाडकर , लक्ष्मीकांत प्यारेलाल , अनिल विश्वास , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल, लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल लता मंगेशकर , मंगेश कदम , सुरेश पंडित , सचिन पिळगावकर , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल , लक्ष्मीकांत प्यारेलाल
महात्मा गांधी	महात्मा गांधी भी अपनी जिंदगी में मर्यादा का पालन करते थे और अगर वे हिंसा का सहारा ले रहे हैं तो वे भी विचारधारा के शिकार हुए हैं। महात्मा गांधी राष्ट्रीय ग्रामीण रोजगार गारंटी योजना के तहत संविदा के कर्मचारियों की हड़ताल का असर इन्दौर में भी दिखा। महात्मा गांधी और उनके समकालीन नेताओं का मुख्य ध्यान समाज के अंतिम व्यक्ति को महत्व देने पर था	महात्मा गांधी जयंती समारोह में भाग लेने के लिए तैयार है। महात्मा गांधी ने कहा है कि उनके पास अपने देश को आजादी के बाद पहली बार राष्ट्रपति बनने की पूरी अधिकारिणी क्षमता है। राष्ट्रपति बनने की उनकी क्षमता है। इस देश को लोकतंत्र का युग मिला है। उन्होंने कहा कि देश में लोकतंत्र की व्यवस्था में लोकतांत्रिक व्यवस्था है। राष्ट्रपति की योग्यता पर विचार करना होगा। महात्मा गांधी ने आज अपने निवास पर एक शोक सभा की अध्यक्षता की ।	महात्मा गांधी को हराते हुए पाकिस्तान ने भी इसे स्वीकार नहीं किया। महात्मा गांधी मैदान में होने वाले कार्यक्रम में बतौर मुख्य अतिथि पहुंचे। महात्मा गांधी की मूर्ति का एलियन को शामिल कर दिया गया	महात्मा गांधी की हत्या से लेकर 1947 से 1947 के बाद भारत की आजादी तक का जो इतिहास और इतिहास के उस पर खड़े होने वाले जो अनेक विमर्श हैं वे भारत में आए हैं । महात्मा गांधी की 150वीं जयंती मनाते हुए भारत सरकार ने स्वतंत्रता आंदोलन में उनके महत्वपूर्ण योगदान के लिए भारत सरकार के राष्ट्रपति भवन, नई दिल्ली से उन्हें पद्मश्री पुरस्कार प्रदान किया। महात्मा गांधी ने एक बार कहा था कि अगर हम गांधी जी के विचारों को अपनाएं तो हम महात्मा गांधी से भी महान हो जाएंगे।	महात्मा गांधी जी को श्रद्धांजलि देने के लिए, एक सप्ताह से भी कम समय बचा है, और तब तक इस मुद्दे के खिलाफ देश भर में कई विरोध प्रदर्शन होने वाले हैं, देश के प्रमुख शहरों में, खासकर, गांधीनगर और वाराणसी में, जहाँ वह एक महात्मा गाँधी हैं और जहाँ, शायद, उनके अंतिम संस्कार के लिए उन्हें ले जाया जा रहा है। महात्मा गांधी , जवाहरलाल नेहरू , इंदिरा गांधी , राजीव गांधी , नरसिंहा राव जैसे नेताओं ने गांधी वादी आंदोलन को खड़ा किया था। महात्मा गांधी के साथ मिलकर चल रहा है

Table 40: Open end text generation: Paramanu-Hindi 162M v/s BLoom (560M, 1.1B, 3B, 7.1B)

Figure 14: Grammar correction task by Paramanu-Bangla, in-context learning capabilities of Paramanu-Bangla model.

Figure 15: Grammar correction task by Paramanu-Bangla-instruct 108.5M, in-context learning capabilities of the model.

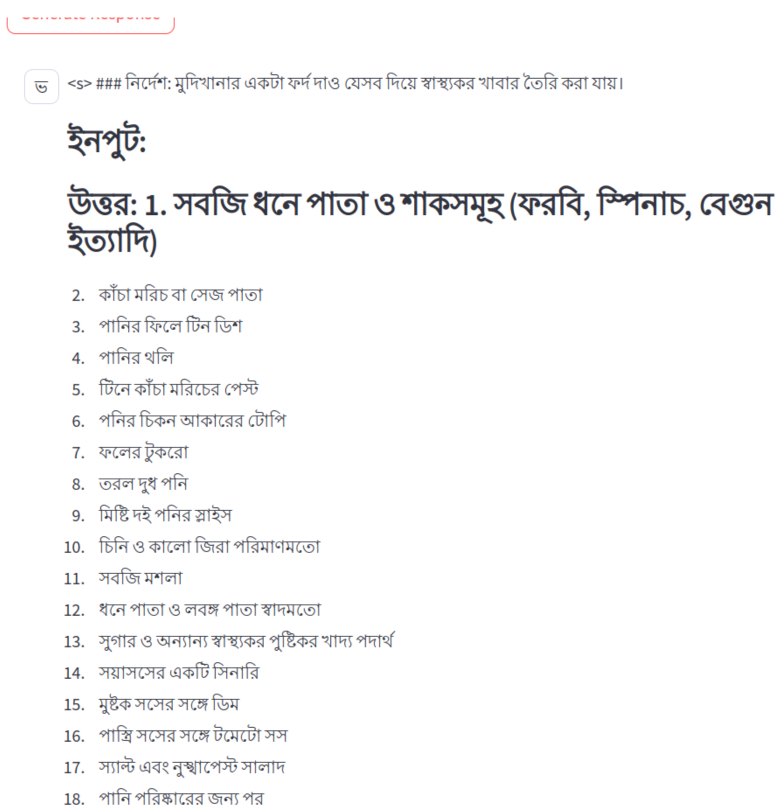


Figure 16: Prompting Paramanu-Bangla-instruct model with query: মুদিখানার একটা ফর্দ দাও যেসব দিয়ে স্বাস্থ্যকর খাবার তৈরি করা যায়। (mudikhānāra ēkaṭā pharda dāō yēsaba diẏē sbāsthyakara khābāra tairi karā yāẏa.)

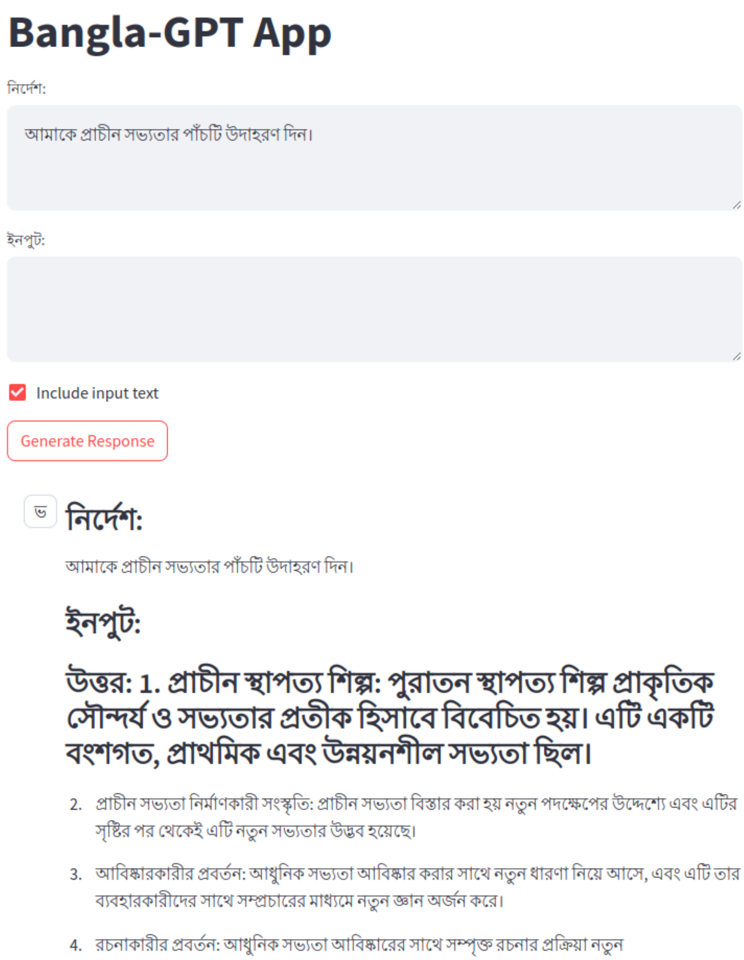


Figure 17: Prompting Paramanu-Bangla-instruct model with query আমাকে প্রাচীন সভ্যতার পাঁচটি উদাহরণ দিন। (āmākē prācīna sabhyatāra pām̐caṭi udāharaṇa dina.)

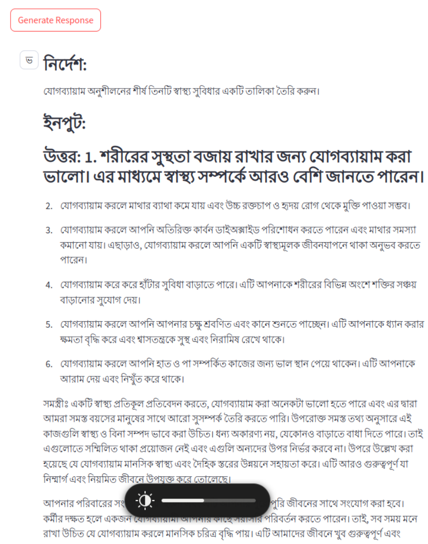


Figure 18: Prompting Paramanu-Bangla-instruct model with query: যোগব্যায়ামের ছয়টি উপকারিতা বলুন। (yōgabyāẏāmēra chaẏaṭi upakāritā baluna.)

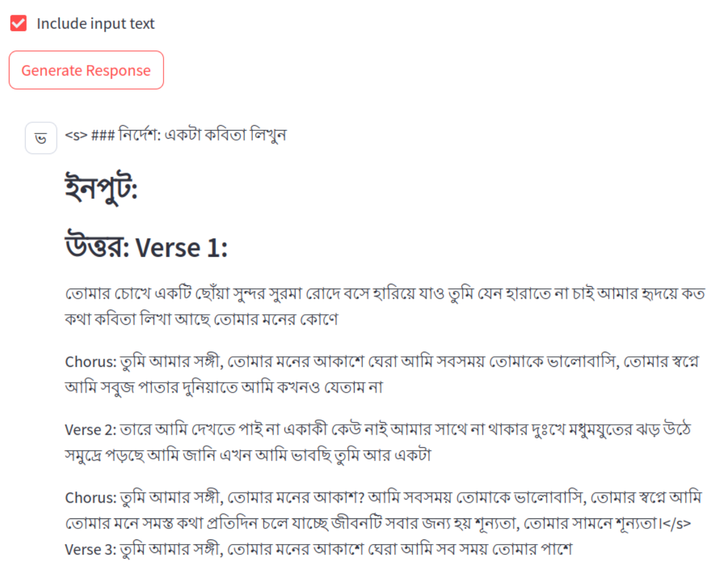


Figure 19: Prompting Paramanu-Bangla-instruct model with query: একটি কবিতা লিখুন। (ēkaṭi kabitā likhuna.)

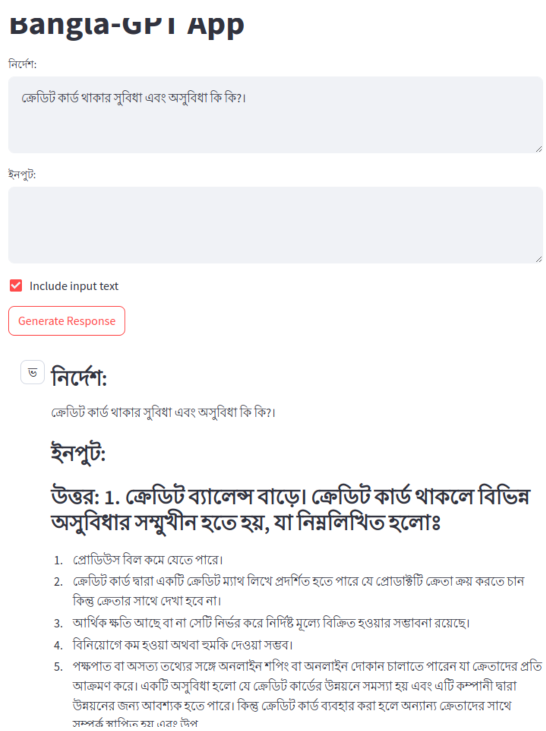


Figure 20: Prompting Paramanu-Bangla-instruct model with query: ক্রেডিট কার্ড থাকার সুবিধা এবং অসুবিধা কি কি? (krēḍiṭa kārḍa thākāra subidhā ēbaṁ asubidhā ki ki?)

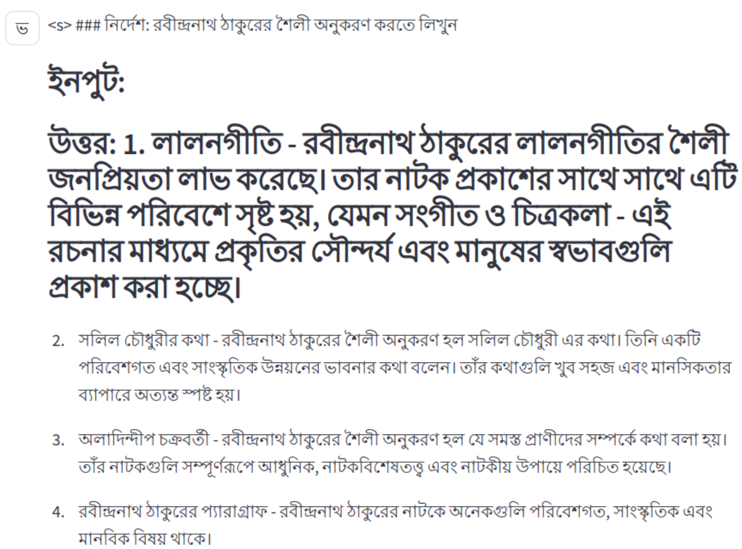


Figure 21: Prompting Paramanu-Bangla-instruct model with query: রবীন্দ্রনাথ ঠাকুরের শৈলী অনুক-রণ করতে লিখুন (rabīndranātha ṭhākurēra śailī anukaraṇa karatē likhuna)

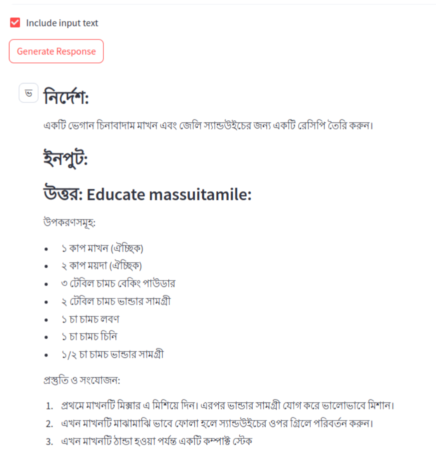


Figure 22: Prompting our instruction tuned Paramanu-Bangla model with query একটি ভেগান চিনাবাদাম মাখন এবং জেলি স্যান্ডউইচের জন্য একটি রেসিপি তৈরি করুন। (ēkaṭi bhēgāna cinābādāma mākhana ēbaṁ jēli syānḍa:uicēra janya ēkaṭi rēsipi tairi karuna.)

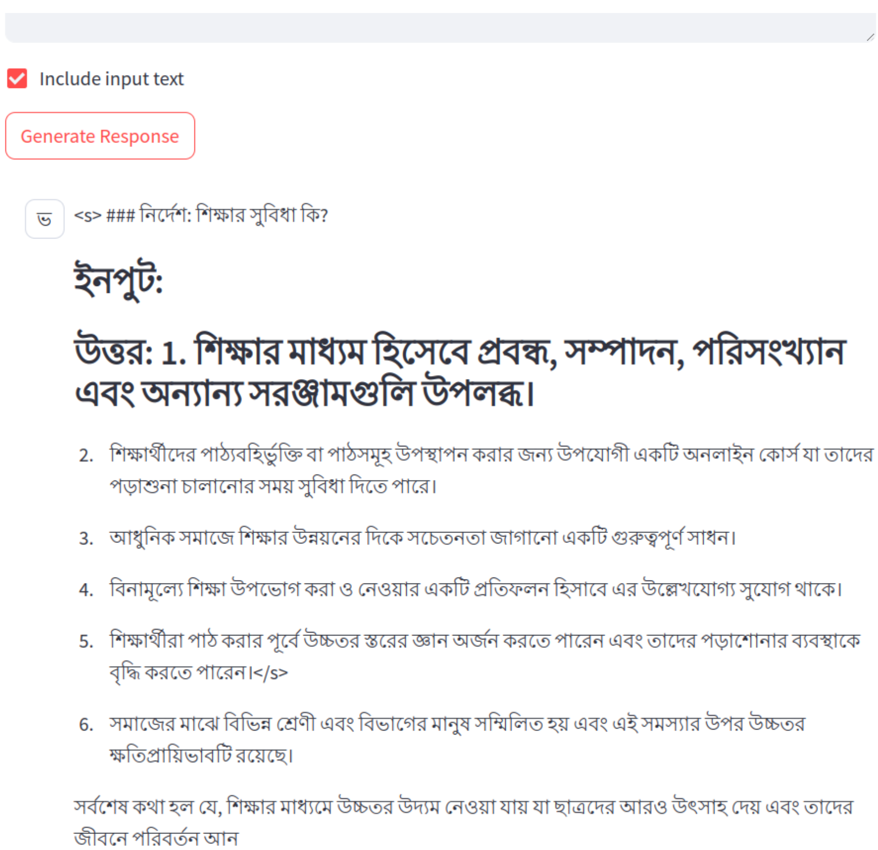
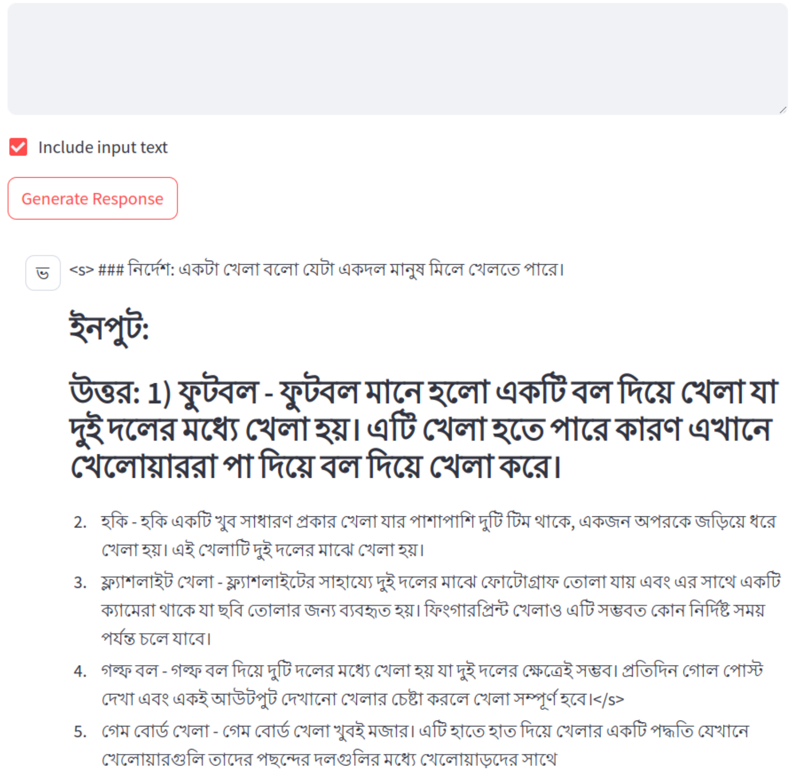


Figure 23: Prompting Paramanu-Bangla-instruct model with query: শিক্ষার সুবিধা কি? (śikṣāra subidhā ki?)

Figure 24: Prompting Paramanu-Bangla-instruct model with query: যৌন শিক্ষার সুবিধা কি? (yauna śikṣāra subidhā ki?)



Include input text

Generate Response

ভ <s> ### নির্দেশ: একটা খেলা বলো যেটা একদল মানুষ মিলে খেলতে পারে।

ইনপুট:

উত্তর: 1) ফুটবল - ফুটবল মানে হলো একটি বল দিয়ে খেলা যা দুই দলের মধ্যে খেলা হয়। এটি খেলা হতে পারে কারণ এখানে খেলোয়াররা পা দিয়ে বল দিয়ে খেলা করে।

2. হকি - হকি একটি খুব সাধারণ প্রকার খেলা যার পাশাপাশি দুটি টিম থাকে, একজন অপরকে জড়িয়ে ধরে খেলা হয়। এই খেলাটি দুই দলের মাঝে খেলা হয়।
3. ফ্ল্যাশলাইট খেলা - ফ্ল্যাশলাইটের সাহায্যে দুই দলের মাঝে ফোটোগ্রাফ তোলা যায় এবং এর সাথে একটি ক্যামেরা থাকে যা ছবি তোলার জন্য ব্যবহৃত হয়। ফিংগারপ্রিন্ট খেলাও এটি সম্ভবত কোন নির্দিষ্ট সময় পর্যন্ত চলে যাবে।
4. গল্ফ বল - গল্ফ বল দিয়ে দুটি দলের মধ্যে খেলা হয় যা দুই দলের ক্ষেত্রেই সম্ভব। প্রতিদিন গোল পোস্ট দেখা এবং একই আউটপুট দেখানো খেলার চেষ্টা করলে খেলা সম্পূর্ণ হবে।</s>
5. গেম বোর্ড খেলা - গেম বোর্ড খেলা খুবই মজার। এটি হাতে হাত দিয়ে খেলার একটি পদ্ধতি যেখানে খেলোয়ারগুলি তাদের পছন্দের দলগুলির মধ্যে খেলোয়াড়দের সাথে

Figure 25: Prompting Paramanu-Bangla-instruct model with query: একটা খেলা বলো যেটা একদল মানুষ মিলে খেলতে পারে। (ēkaṭā khēlā balō yēṭā ēkadala mānuṣa milē khēlatē pārē.)

ভ <s> ### নির্দেশ: একটা গল্প লিখুন

ইনপুট:

উত্তর: "আমি একটি ছোট মেয়ে। আমার জীবন পরিবর্তিত হয়ে গেছে। আমি কখনো খুশি হতে পারি না আর কখনও হবে না। কিন্তু আসলে একটা মজার জিনিস আছে যা আমাকে বেশ আনন্দ দেয়। তাই আমি ঘরে থাকতেই ভালোবাসি নতুন কিছু।

আমার জীবনের প্রথমদিনগুলোতে আমি একটি অদ্ভুত অভিজ্ঞতা পেয়েছি। আমি ভাবতে পারছি যে মানুষ কখনোই ভুল করেনা ভুল করে এমন কথাগুলো বললে তখন তার সাথে থাকতে কত সুবিধা হবে। আর সেই সময়টাকে আমি কখনো ভুলবো না। এটি আমার জীবনের একটা অসাধারণ অভিজ্ঞতা ছিল।

যখন আমি স্কুলে পড়তাম, আমি সব সময় খুব খুশি হতে পারতাম এবং আমার শিক্ষার্থীদের উন্নতি করতে হতো। স্কুলের ক্লাসে সবাই আমাকে দেখতে আসত হাসাতে হাসতে। কিন্তু স্কুল থেকে চলে এসে একটি দিনগুলো আমার জীবনে ভালো লেগে গেল। আমি একটা ছাতা নিয়ে স্কুলে উঠে গেলাম এবং স্কুলের ফার্নিচার কিনে ফিরে এসে দেখি একটা বাথরুমের দরজা খোলা। সেখানে আমি পরিষ্কার ভাবে খাবার বানিয়ে নিলাম এবং সবাইকে উপহার দিলাম।

শুরুতেই আমি সকল বন্ধুদের কথা শুনে আসলাম কিন্তু সেই ছোট ঘটনা আমার মনেও কখনোই শান্তি আনবে না। আমি ভাবছিলাম আমার বয়স আরো বাড়বে কিনা, কিভাবে আমি এত কিছু হারানো হবো! কিন্তু তখন স্কুলের পরিবেশ খুব বিচিত্র ছিল। সেখানে সবাই মজার মজার কথা শেয়ার করছিল। তাদের হাসি আমার মুখে লালচে ফুটে যাচ্ছিল এবং একটি সুন্দর স্মৃতি আমার হৃদয়ে জন্ম নেওয়া হয়েছিলো।

এখন এই বিষয়টি মনে পড়ে আমি স্কুলে যাচ্ছি না, তবে আমি যখন স্কুলের পরিবেশ সম্পর্কে ভাবতে শুরু করলাম, তখন আমার মনে হলো যে, এই রকম একটা স্থিতি সবার জন্য সুখ, সম্মান এবং প্রতিষ্ঠার ক্ষেত্রে। এটা সম্ভব যদি আমি নতুন কিছু শিখে যাই, তাহলে আমার জীবনে একটা পরিবর্তন আসতে পারে।

Figure 26: Paramanu-Bangla 108.5M generation capabilities at maximum tokens limit of 1024 for story generation in Bangla. Query: একটি দীর্ঘ গল্প লিখুন (ēkaṭi dīrgha galpa likhuna)

আমার স্কুলে পরিবেশ সম্পর্কে আমরা শেখা শিখলাম, যদিও সেই স্কুলে থাকতে খুব একটা সহায়তা ছিলো না। কিন্তু আমরা যখন বাসায় খেলতাম, তখন আমাদের মনে হতো নতুন কিছু শিখেছি। আমি এখন জানি সেটার উপর নির্ভর করে অনেক কঠিন টেকনোলজি দরকার হবে। আমার স্কুলে থাকতে থাকতে মনে হতে থাকে আমাদের জীবনে সমস্যা আর একটি স্কুলে থাকতে হলে সেই স্কুলাতে নিজেকে সাম্প্রতিক উপাদানগুলি বিনিময় করা উচিত। তারপরও, আমি নিশ্চিত হই যে আমরা সমস্যার সমাধানে সাফল্য অর্জন করতে পারব।

পরিবেশ সম্পর্কে কথা বলতে গেলে আমি জানতে পেরেছি যে আমাদের সবচেয়ে গুরুত্বপূর্ণ জিনিস হলো নিজেদের সুস্থ রাখা। নিজের শরীর স্বাস্থ্য পরিচালনার জন্য একটি কম্পিউটারের প্রয়োজন হয়, যাতে তার সব উপকরণ স্বাস্থ্যকর থাকে। পরিবেশের উপর কন্ট্রোলের মাধ্যমে নির্ভুল কাজ করার ব্যবস্থা নিয়ে শিক্ষা নেওয়ার চেষ্টা করা উচিত। শিক্ষণ প্রণালী একটি আদর্শ পদক্ষেপ যা দক্ষতা ও প্রযুক্তি উন্নয়নে আমাদের সাহায্য করবে।

পরবর্তীতে, আমাদের বিভিন্ন শিক্ষা পদ্ধতি একত্রিত করে একটি সমর্থনকারী দল গঠিত হয়েছে যারা বিভিন্ন প্রকল্পে অংশগ্রহণ করেছে। উদাহরণস্বরূপ, আমাদের স্কুলের পড়াশোনার জন্য সমস্ত বিবরণগুলি সংগ্রহ করা হয়েছে। তারপর একটি সম্মেলনের সময়, শিক্ষকরা ছাত্র-ছাত্রীদের মধ্যে কথোপকথন চালু করে দিয়েছেন এবং সবাইকে একটি নির্দিষ্ট পরিমাণের সম্মতি দিতে বলেছেন। শিক্ষকরা এক্সপান্ড করেছেন একটি স্থায়ী সম্মতি।

পরিবেশে পরিবর্তন এবং সামাজিক উন্নয়নের উদ্দেশ্যে এই সমর্থনকারীরা ছাত্র-ছাত্রীদের পাঠশালায় উপযোগী পাঠ প্রদর্শন করেছেন। এটি শুভ হোক এবং শুভ হোক।</s> পরের পরিকল্পনাটি স্কুলে শিক্ষকদের নেতৃত্বে পরীক্ষা দেওয়া এবং সমস্যার সমাধান করা। শিক্ষকরা ছাত্র-ছাত্রীদের কাছে একটি সম্পূর্ণ শিক্ষামূলক অনুভূতি দেওয়ার সুযোগ প্রদান করে এবং তারা সমস্যার সমাধান করার জন্য সর্বোচ্চ মানসম্পন্ন হন।

এই পদ্ধতি সম্পূর্ণ নতুন করে কিছু সমস্যাগুলি সমাধান করা হয় এবং আমরা আগের চেয়ে অনেক বেশি চেষ্টা করি যাতে কোনো সমস্যা না হয়।</s> প্রথমবারের মতো, আমার মতো মানুষদের জন্য এই একটি সমর্থনকারীর গল্পে পরিচিত হওয়া গল্পটি শুরু হল। এখানে একটি প্রশ্ন উঠে এসেছে এবং প্রশ্নটি ছিল - কিভাবে একটি স্কুলে শিক্ষক এবং শিক্ষার্থীরা একটি সমস্যার সমাধান করতে পারে? এই গল্পে, আমরা জানি না যে কীভাবে শিক্ষক এবং শিক্ষার্থীদের পেছনে পিছনে পথ চলা যায়।</s> পরবর্তী প্রশ্ন হল - এই গল্পটি কী ভাবে আমাদের সমস্যা সমাধান করে? এখানে একটি

Figure 27: Long story generation, continued

Include input text

Generate Response

ভ <s> ### নির্দেশ: কেন মানুষ ঈশ্বরে বিশ্বাস করতে বিকশিত হয়েছে?

ইনপুট:

উত্তর: 1. সমস্ত ধর্মের মূল হল ঈশ্বর ও পরকালের বিশ্বাস। এটি পরম ঈশ্বরের সাক্ষাৎ অনুভব, জীবন ও মৃত্যুর উপর বিশেষ জোর দেয়।

2. ঈশ্বরের বিচারে নির্বিচার ও অসত্য বিচার করা একটি জাতিগত পরিণতি যা প্রতিবেদনকারীদের দ্বারা প্রতিপাদিত হয়। এতে সমস্ত ধার্মিক এবং নৈতিক মানদণ্ডগুলি রয়েছে, যা নির্বিচার ও অসত্যকে প্রশংসা করে।

3. ঈশ্বরের সৃষ্টি একটি অদ্বিতীয় সৃষ্টি যা আমাদের জীবনে কখনও ঘটে না। এটি আমাদের সৃষ্টিশীলতার ও আমাদের জীবনের লক্ষ্যে বিশেষ গুরুত্ব রাখে।

4. ঈশ্বরের শাস্ত্রীয় তত্ত্ব একটি বিশ্বজগতে প্রদর্শিত হয় যা সমস্ত ধারণাকে নিয়ন্ত্রণ করে। এটি আমরা দেখতে পাই আমাদের সমাজে এবং সব গ্রহের পরিস্থিতিতে যখন ঈশ্বরের প্রভাব অস্বাভাবিক লাগে তখন তা অবিচলিত থাকে।

সুতরাং, মানুষ ঈশ্বরে বিশ্বাস করলে সেটি তার পরিবর্তনকারী হিসাবে প্রকাশ পায় এবং তার পরিবর্তনে ঈশ্বর আমাদের কাছে একটি গুরুত্বপূর্ণ ভূমিকা পালন করে।</s> তাই মানুষের

Figure 28: Prompting Paramanu-Bangla-instruct model with query: কেন মানুষ ঈশ্বরে বিশ্বাস করতে অনুপ্রাণিত হয়েছে? (kēna mānuṣa īśbarē biśbāsa karatē anuprāṇita haẏēchē?)

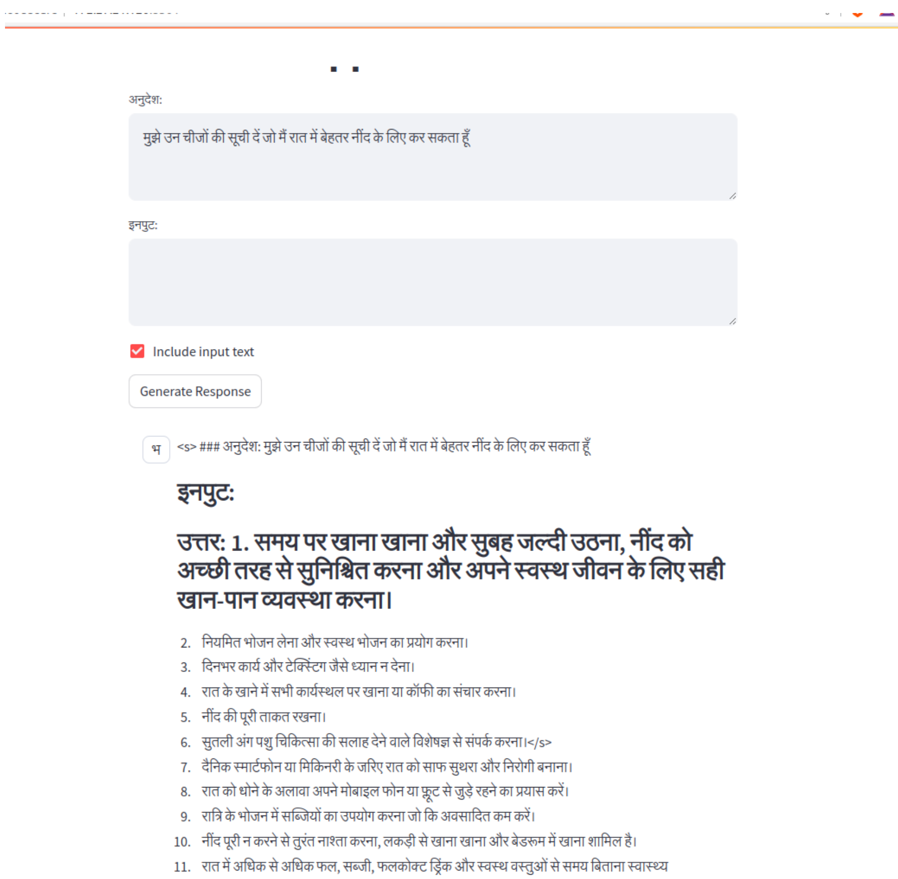


Figure 29: Prompting Paramanu-Hindi-instruct 367.5M model with query: मुझे उन चीजों की सूची दें जो मैं रात में बेहतर नींद के लिए कर सकता हूँ (mujhē una cījōṁ kī sūcī dēṁ jō maiṁ rāta mēṁ bēhatara nīṁda kē liē kara sakatā hūm̐)

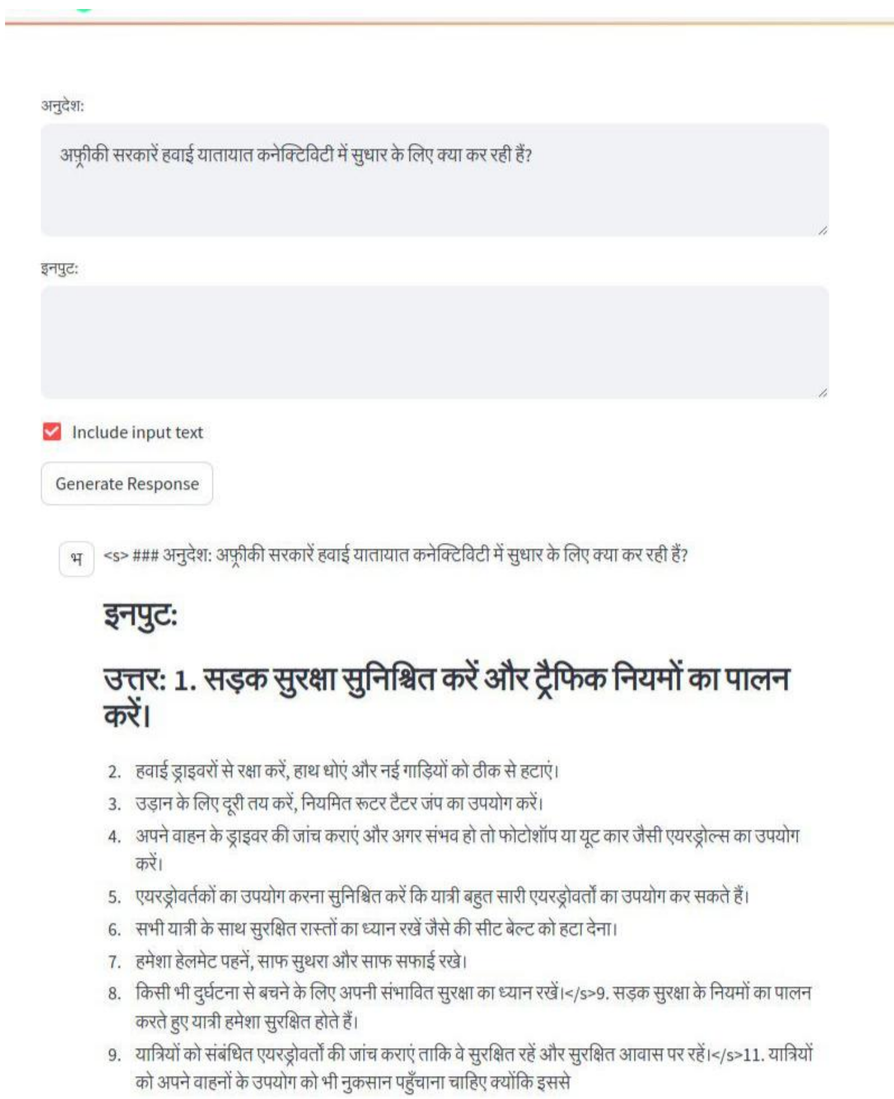


Figure 30: Prompting Paramanu-Hindi-instruct 367.5M model with query: अफ़्रीकी सरकारें हवाई यातायात कनेक्टिविटी में सुधार के लिए क्या कर रही हैं? (afrīkī sarakārēṁ havāī yātāyāta kanēkṭiviṭī mēṁ sudhāra kē liē kyā kara rahī haiṁ?)

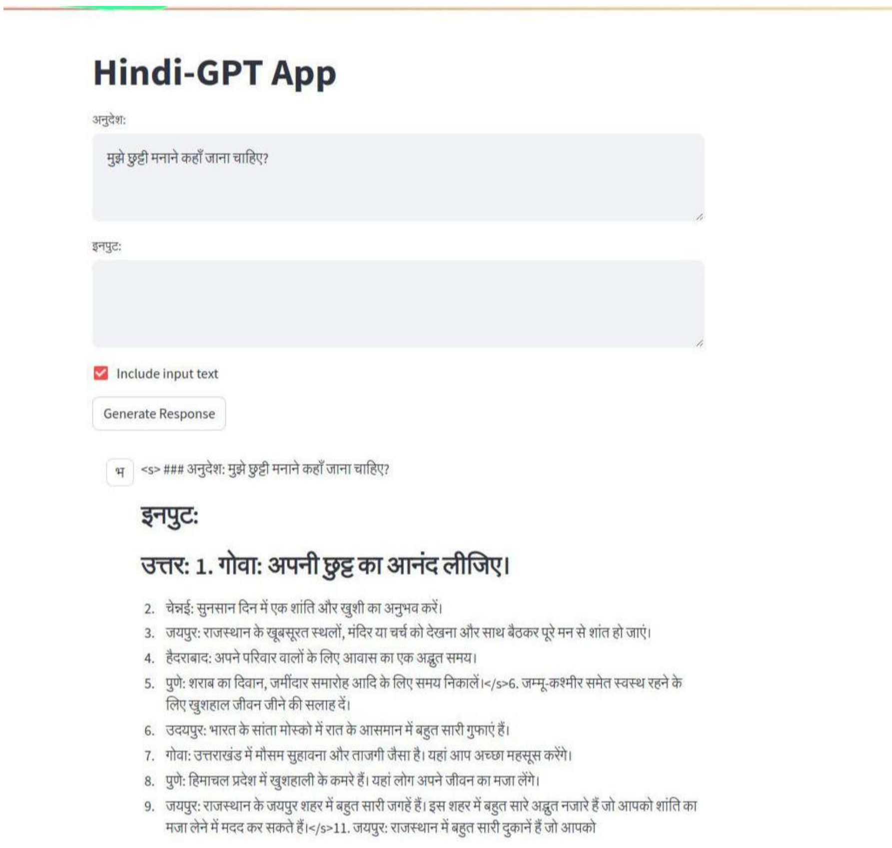


Figure 31: Prompting Paramanu-Hindi-instruct 367.5M model with query: मुझे छुट्टी मनाने कहाँ जाना चाहिए? (mujhē chuṭṭī manānē kahāṁ̐ jānā cāhiē?)

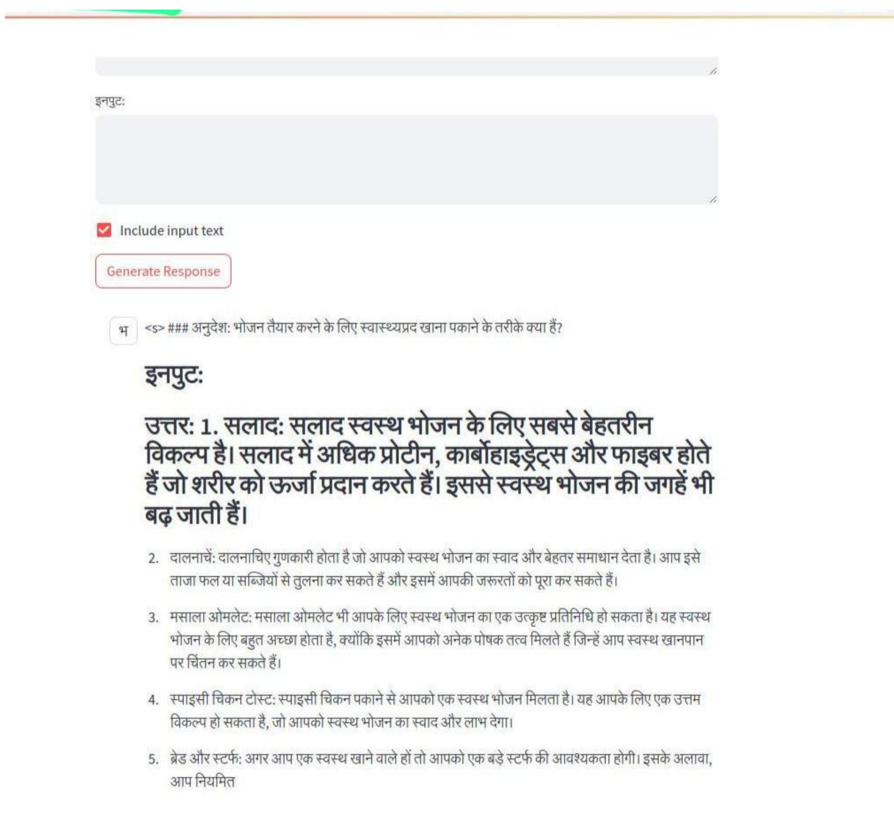


Figure 32: Prompting Paramanu-Hindi-instruct 367.5M model with query: भोजन तैयार करने के लिए स्वास्थ्यप्रद खाना पकाने के तरीके क्या हैं? (bhōjana taiyāra karanē kē liē svāsthyaprada khānā pakānē kē tarīkē kyā haiṁ?)

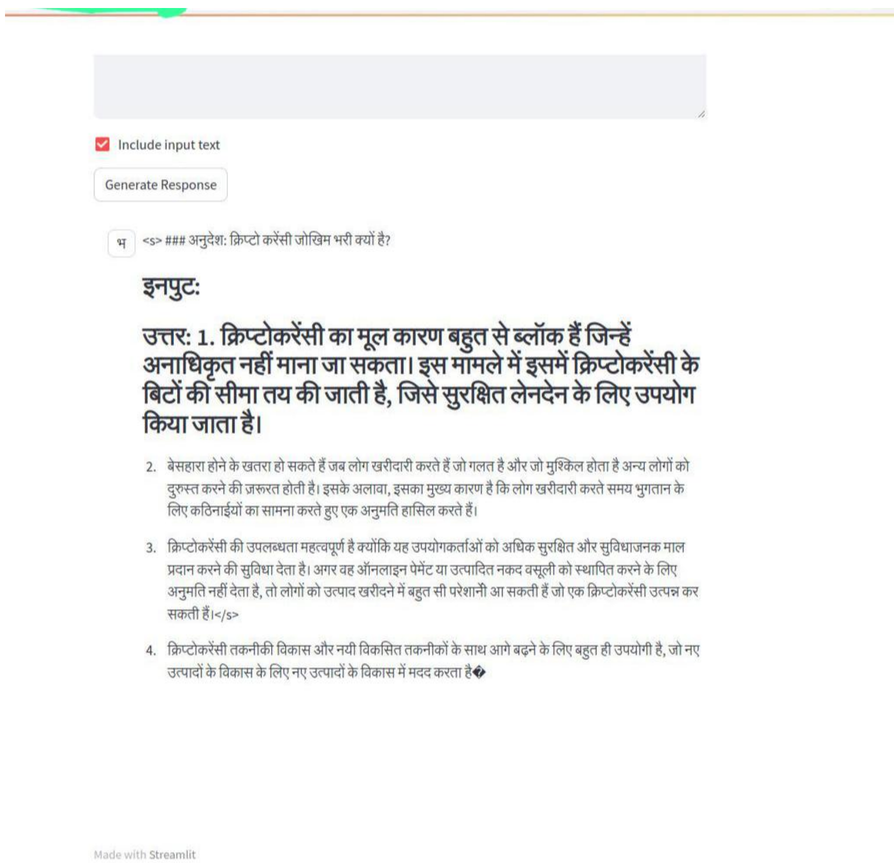


Figure 33: Prompting Paramanu-Hindi-instruct 367.5M model with query: क्रिप्टो करेंसी जोखिम भरी क्यों है? (kripṭō karēṁsī jōkhima bharī kyōṁ hai?)

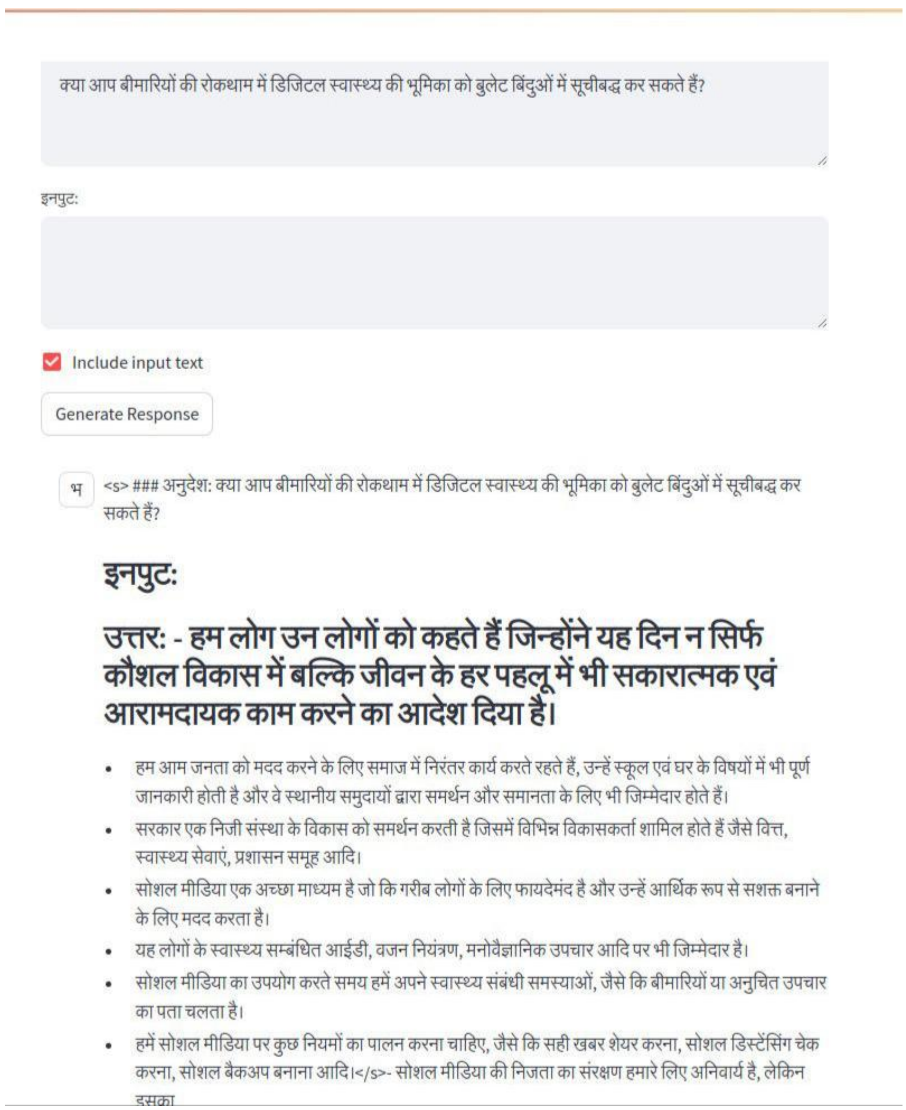


Figure 34: Prompting Paramanu-Hindi-instruct 367.5M model with query: क्या आप बीमारियों की रोकथाम में डिजिटल स्वास्थ्य की भूमिका को बुलेट बिंदुओं में सूचीबद्ध कर सकते हैं? (kyā āpa bīmāriyōṁ kī rōkathāma mēṁ ḍijiṭala svāsthya kī bhūmikā kō bulēṭa biṁduōṁ mēṁ sūcībaddha kara sakatē haiṁ?)

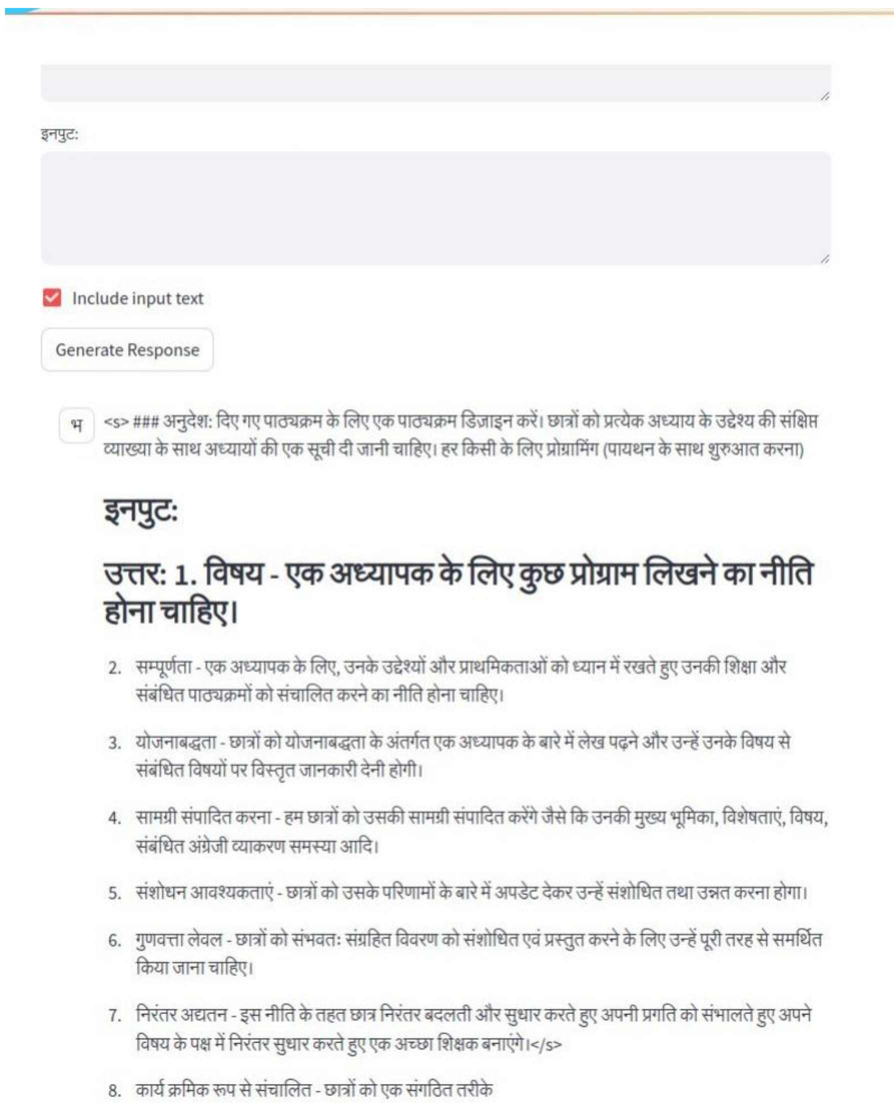


Figure 35: Prompting Paramanu-Hindi-instruct 367.5M model with query: दिए गए पाठ्यक्रम के लिए एक पाठ्यक्रम डिज़ाइन करें। छात्रों को प्रत्येक अध्याय के उद्देश्य की संक्षिप्त व्याख्या के साथ अध्यायों की एक सूची दी जानी चाहिए। हर किसी के लिए प्रोग्रामिंग (पायथन के साथ शुरुआत करना) (diē gaē pāṭhyakrama kē liē ēka pāṭhyakrama ḍizāina karēṁ. chātrōṁ kō pratyēka adhyāya kē uddēśya kī saṁkṣipta vyākhyā kē sātha adhyāyōṁ kī ēka sūcī dī jānī cāhiē. hara kisī kē liē prōgrāmiṁga (pāyathana kē sātha śuruāta karanā))

Figure 36: Prompting Paramanu-Hindi-instruct 367.5M model with query: 1970 और 1980 के दशक के दौरान अधिकांश अफ्रीकी देशों पर सैन्य तानाशाही का नियंत्रण क्यों था? 1970 aura 1980 (kē daśaka kē daurāna adhikāṁśa aphrīkī dēśōṁ para sainya tānāśāhī kā niyaṁtraṇa kyōṁ thā?)

भ <s> ### अनुदेश: प्रेम क्या है?

इनपुट:

उत्तर: । प्रेम एक भाव होता है जो दूसरे का प्रतिबिम्ब बनाता है और स्नेह की भावना दर्शाता है। यह विश्वास, संवेदना और अंतर्निहित भावनाएं, जब आप महसूस करते हैं कि आपकी इच्छा किसी भी व्यक्ति या स्थिति के लिए स्थिर नहीं होती हैं, आपके जीवन में एक उज्ज्वल स्वर्णिम धारा बनता है। प्रेम एक भावना का रूप हो सकता है, जो आदर्शों से भरा हुआ होता है जो एक दूसरे को मदद करने में सक्षम होते हैं।</s></s>प्रेम एक ऐसी भावना होती है जिसमें हम स्वयं अपना समय, ऊर्जा या अनुभूति समझते हुए अन्य लोगों की आवाजों, जातिवाद, लिंग अथवा नास्तकों आदि से जुड़ जाते हैं। इस भावना को याद रखना और उसे अपने जीवन में लागू करना हमारी जिम्मेदारी है। इसलिए, प्रेम एक महत्वपूर्ण मानवीय सामाजिक वातावरण है जो हमें अपने समाज में एक बेहतर, संतुष्ट और संतुलित नागरिक बनाती है। </s></s>प्रेम एक जीवन का आधार और एक महत्वपूर्ण अंग होता है, जहां हर इंसान कहीं-कहीं एकता के लिए जिम्मेदार होता है, जहाँ सच्ची भावनाएँ नहीं होतीं हैं। इसलिए, प्रेम हमारे जीवन का सबसे मूल आधार है।</s>

Figure 37: Prompting Paramanu-Hindi-instruct 367.5M model with query: प्रेम क्या है? (prēma kyā hai?)

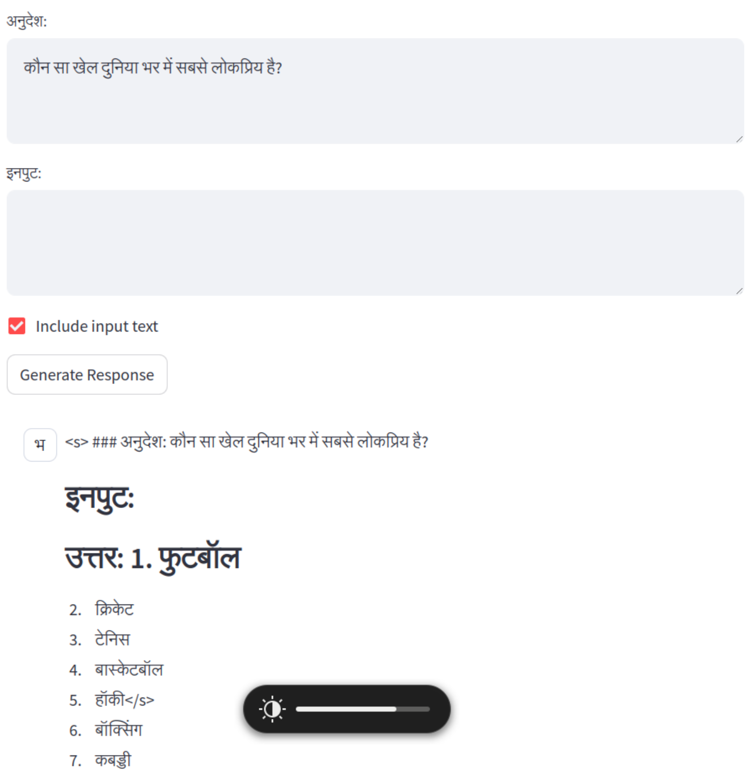


Figure 38: Prompting Paramanu-Hindi-instruct 367.5M model with query: कौन सा खेल दुनिया भर में सबसे लोकप्रिय है? (kauna sā khēla duniyā bhara mēṁ sabasē lōkapriya hai?)

Figure 39: Prompting Paramanu-Tamil-instruct model with query: கல்லூரிப் பட்டப்படிப்பைத் தொடர்வதன் 5 நன்மைகளைப் பட்டியலிடுங்கள். (kallūrip paṭṭappaṭippait toṭarvataṉ 5 naṉmaikaḷaip paṭṭiyaliṭuṅkaḷ.)

ப

அறிவுறுத்தல்:

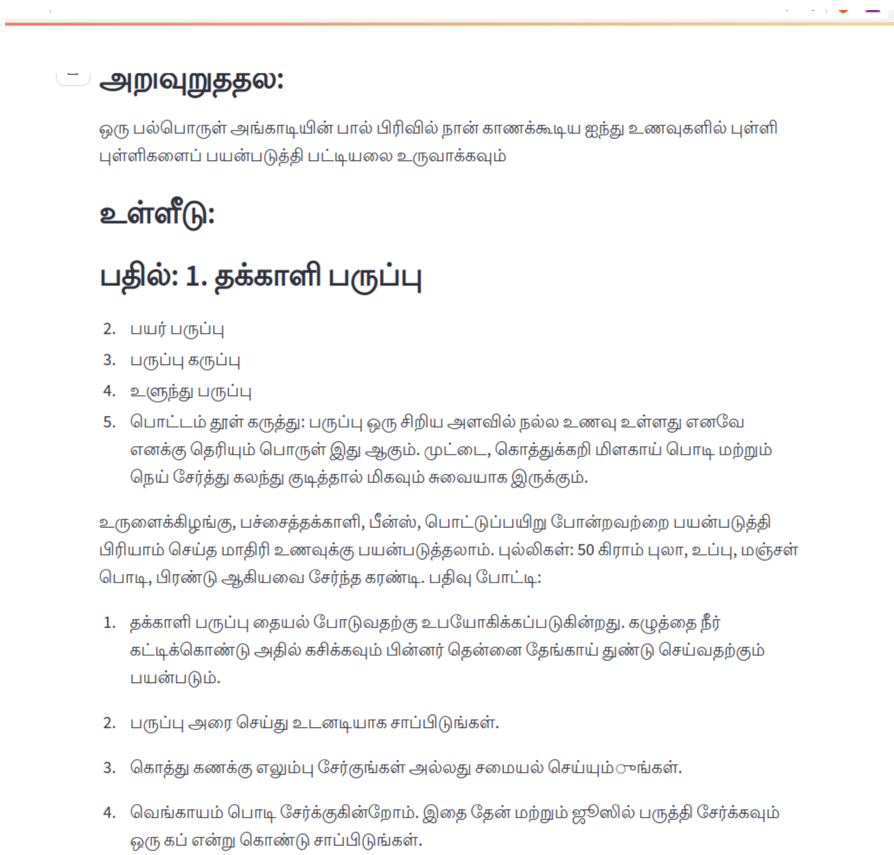
புவிசார் தரவு என்றால் என்ன?

உள்ளீடு:

பதில்: ◌ புரம் புவிசார் தரவு என்பது பூமியின் அடிப்படை தரவுகளை வழங்கும் வழிகளை வழங்கும் கணினி நிரல் என்று குறிப்பிடுகின்றது. இது புவியின் மொத்த இயற்கை நிலங்கள் மற்றும் உயிரினங்களை பற்றி ஆராய்ந்து விளக்கப்படுகின்றது. இதில் உள்ள கணினி நிரல்கள் பல பிரபல நிரல்களின் பயன்பாடுகளைக் கொண்டு வருகின்றன. இவைகளில் சிலவேனும் பூ மற்றும் காய்கறிகளுடன் தொடர்பு கொண்டுள்ளதாக இருக்கின்றன.

புவிசார் தரவு போன்றது பூவின் அடிப்படை தரவுகளைளிக்கும் பயனர் நிரல்களின் மூலம் மட்டுமே வழங்கப்படுகின்றது. இது புவியின் பெரும்மையான பிரிவுகள் மற்றும் உயிரினங்களை ஆராய்ந்து அறியலாம். இது மிகவும் பயனுள்ள ஒரு செயல் ஆகும் மேலும் மக்கள் மற்றும் நிரல் இயக்கக் குழுக்களுக்கு பயன்படுத்தி சிறகுவதுமான பலவீனத்தை பெறும்.புவிசார் தரவு இவ்வாறு எவ்வாறு பயன்படுத்தப்படுகின்றது என்பதை புரிந்துகொள்ள இதைarltamil.com இல் தொடர்பு கொள்ளவும்.

Figure 40: Prompting Paramanu-Tamil-instruct model with query: புவிசார் தரவு என்றால் என்ன? (puvicār taravu eṉṟāl eṉṉa?)



அறாவுறுததல:

ஒரு பல்பொருள் அங்காடியின் பால் பிரிவில் நான் காணக்கூடிய ஐந்து உணவுகளில் புள்ளி புள்ளிகளைப் பயன்படுத்தி பட்டியலை உருவாக்கவும்

உள்ளீடு:

பதில்: 1. தக்காளி பருப்பு

2. பயர் பருப்பு
3. பருப்பு கருப்பு
4. உளுந்து பருப்பு
5. பொட்டம் தூள் கருத்து: பருப்பு ஒரு சிறிய அளவில் நல்ல உணவு உள்ளது எனவே எனக்கு தெரியும் பொருள் இது ஆகும். முட்டை, கொத்துக்கறி மிளகாய் பொடி மற்றும் நெய் சேர்த்து கலந்து குடித்தால் மிகவும் சுவையாக இருக்கும்.

உருளைக்கிழங்கு, பச்சைத்தக்காளி, பீன்ஸ், பொட்டுப்பயிறு போன்றவற்றை பயன்படுத்தி பிரியாம் செய்த மாதிரி உணவுக்கு பயன்படுத்தலாம். புல்லிகள்: 50 கிராம் புலா, உப்பு, மஞ்சள் பொடி, பிரண்டு ஆகியவை சேர்ந்த கரண்டி. பதிவு போட்டி:

1. தக்காளி பருப்பு தையல் போடுவதற்கு உபயோகிக்கப்படுகின்றது. கழுத்தை நீர் கட்டிக்கொண்டு அதில் கசிக்கவும் பின்னர் தென்னை தேங்காய் துண்டு செய்வதற்கும் பயன்படும்.
2. பருப்பு அரை செய்து உடனடியாக சாப்பிடுங்கள்.
3. கொத்து கணக்கு எலும்பு சேர்குங்கள் அல்லது சமையல் செய்யும்ுங்கள்.
4. வெங்காயம் பொடி சேர்க்குகின்றோம். இதை தேன் மற்றும் ஜூஸில் பருத்தி சேர்க்கவும் ஒரு கப் என்று கொண்டு சாப்பிடுங்கள்.

Figure 41: Prompting Paramanu-Tamil-instruct model with query:ஒரு பல்பொருள் அங்காடியின் பால் பிரிவில் நான் காணக்கூடிய ஐந்து உணவுகளில் புள்ளி புள்ளிகளைப் பயன்படுத்தி பட்டியலை உருவாக்கவும் (oru palporuḷ aṅkāṭiyiṉ pāl pirivil nāṉ kāṇakkūṭiya aintu uṇavukaḷil puḷḷi puḷḷikaḷaip payaṉpaṭutti paṭṭiyalai uruvākkavum)

உள்ளீடு:

பதில்: ◌ாம் நாட்டின் கருத்து மக்களிடம் தேசிய பூங்காக்களை விரும்புகின்றன. இந்த தேசிய பூங்காக்கள் மக்களுக்கு மிகவும் முக்கியமானவைகளாக உள்ளனவும் அவற்றுக்கு மக்கள் தங்க மூலம் முன்பதிவு செய்யப்படுகின்றனர். இது மக்கள் மதிப்பில் தேசியபூங்காக்களின் மதிப்புகளை அறிய உதவுகின்றது. மேலும், மக்கள் தங்கள் தேசிய பூங்காக்களுக்குத் தேவையான நிகழ்வுகளை தெரிந்து கொள்ள உதவுகின்றனர். இது மக்களின் மென்மையை உயர்த்துகின்றது மற்றும் அக்கறைக்கு உதவுகின்றது.

ஒரு பூங்காக்களில் புதிய கருத்துக்கள் மற்றும் பயிற்சிகள் இருக்கும். இவை மக்கள் தேசிய பூங்காக்களில் வருகின்ற உற்பத்திகளை விட மிகவும் முக்கியமானவை என கருதப்படுகின்றது. இது மக்களின் கருத்துக்களை மேம்படுத்துகின்றது மற்றும் அவர்களின் வாழ்க்கைக்கு உதவுகின்றது. இது மக்கள் தேசிய பூங்காக்களுக்குத் தேவையான நிகழ்வுகளை மட்டும் அல்லது அவர்கள்ent◌ிகை பள்ளிகளுக்கு உதவுகின்றது. இது மக்களின் உணர்வுகளை மேம்படுத்துகின்றது மற்றும் அவர்கள் எப்படி தேசிய பூங்காக்களுக்கு உதவ முடியும் என்று தெரியும்.அதனால் மக்கள் தேசிய பூங்காக்களை காண வேண்டும் மற்றும் அதன் பின்புறம் இருந்துாங்க வேண்டும். இது அவர்களின் வாழ்க்கைக்கு உதவுகின்றது மற்றும் அவர்களின் பகிர்வுகளுக்கு மேற்பட்ட வாய்ப்புகளை வழங்குகின்றன.பாரா தகுதி ஆனது தேசிய பூங்காக்கள்! தேசிய பூங்காக்கள் மிகவும் முக்கியமாக உள்ளன. இது மக்களின் உயர்வுகளை போக்குகிறது மற்றும் அவர்களின் வாழ்வில் உதவுகின்றது. எனவே, மக்கள்

Figure 42: Prompting Paramanu-Tamil-instruct model with query:குடியிருப்பாளர்களுக்கு தேசிய பூங்காவின் சில நன்மைகள் என்ன?? (kuṭiyiruppāḷarkaḷukku tēciya pūṅkāviṉ cila naṉmaikaḷ eṉṉa?)

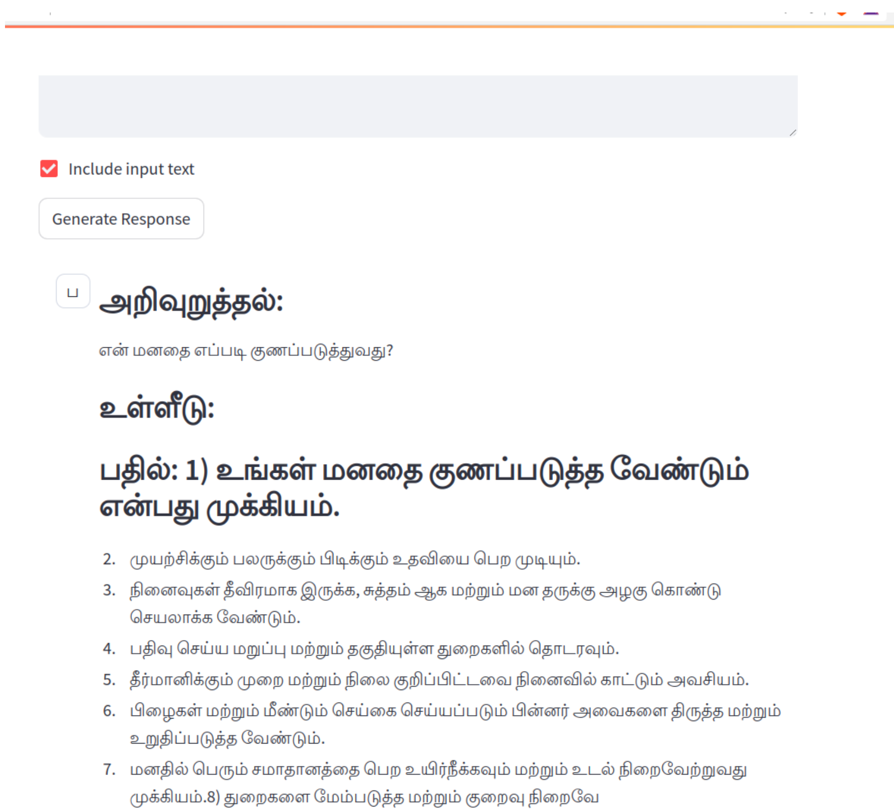


Figure 43: Prompting Paramanu-Tamil-instruct model with query: என் உணர்வுகளை எப்படி குணப்படுத்துவது? (eṉ uṇarvukaḷai eppaṭi kuṇappaṭuttuvatu?)

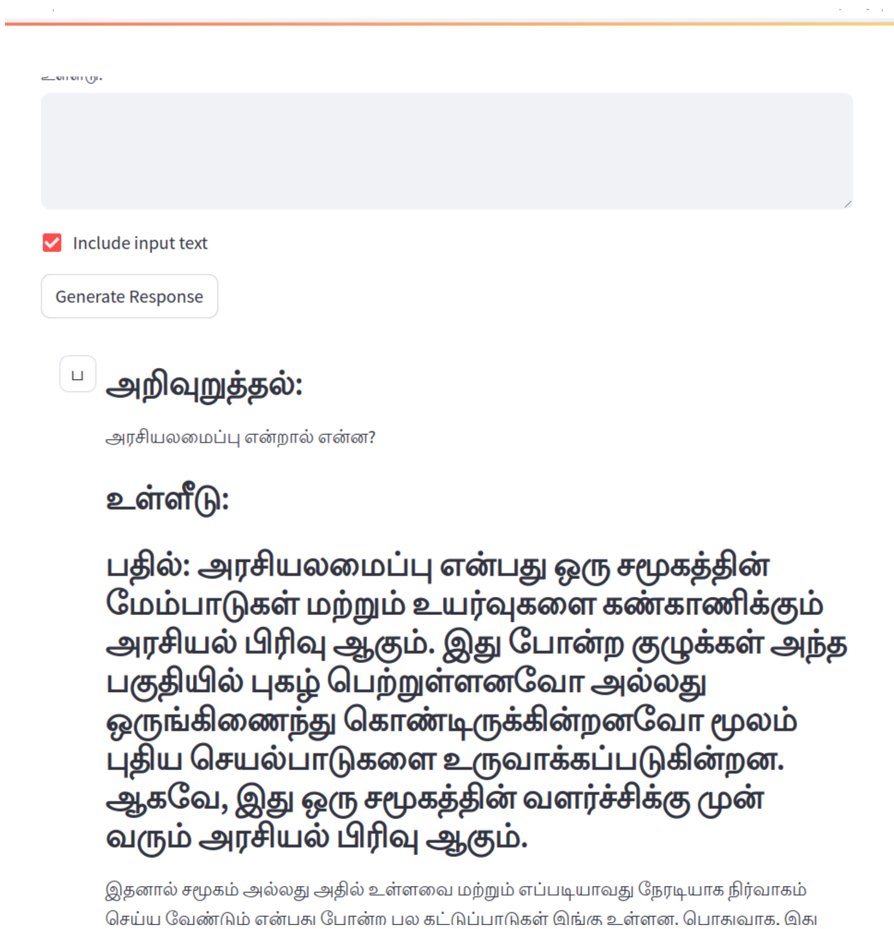


Figure 44: Prompting Paramanu-Tamil-instruct model with query: அரசியலமைப்பு என்றால் என்ன? (araciyalamaippu eṉṟāl eṉṉa?)

Figure 45: Prompting Paramanu-Telugu-instruct model with query: అగ్నిపర్వతాలు ఎలా ఏర్పడతాయి? (agniparvatālu elā ērpaḍatāyi?)

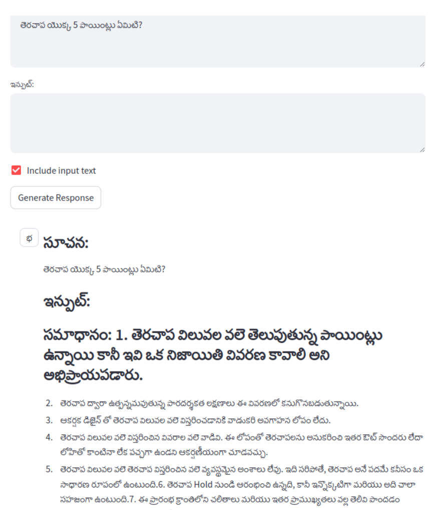


Figure 46: Prompting Paramanu-Telugu-instruct model with query: తెరచాప యొక్క 5 పాయింట్లు ఏమిటి? (teracāpa yokka 5 pāyiṁṭlu ēmiṭi?)

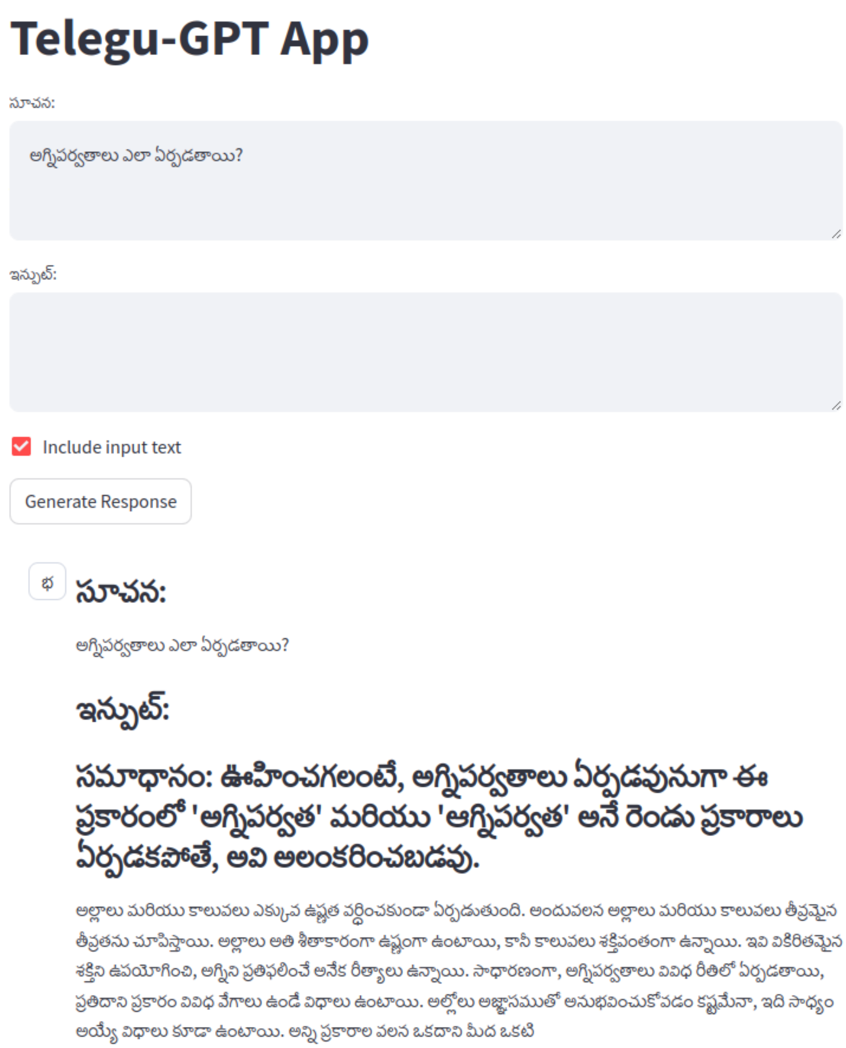


Figure 47: Prompting Paramanu-Telugu-instruct model with query: అగ్నిపర్వతాలు ఎలా ఏర్పడతాయి? (agniparvatālu elā ērpaḍatāyi?)

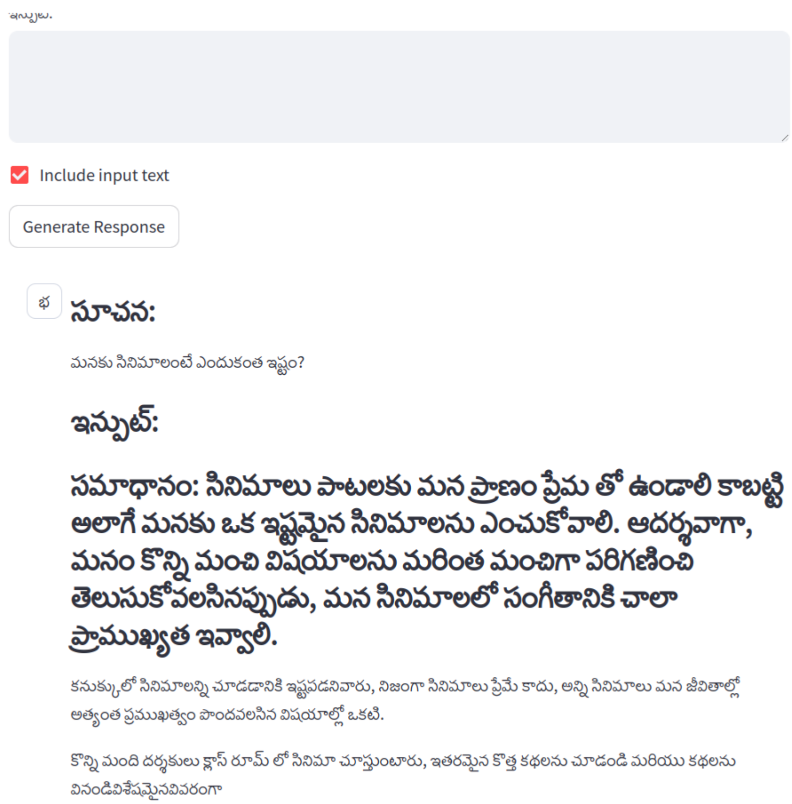


Figure 48: Prompting Paramanu-Telugu-instruct model with query: మనకు సినిమాలంటే ఎందుకంత ఇష్టం? (manaku sinimālaṁṭē eṁdukaṁta iṣṭaṁ?)

Figure 49: Prompting Paramanu-Telugu-instruct model with query: చిరంజీవి (ciraṁjīvi)

You may include other additional sections here.