

BenLLM-Eval: A Comprehensive Evaluation into the Potentials and Pitfalls of Large Language Models on Bengali NLP

Mohsinul Kabir^{1,*}, Mohammed Saidul Islam^{2,*}, Md Tahmid Rahman Laskar^{2,5,†},
Mir Tafseer Nayeem³, M Saiful Bari⁴, Enamul Hoque²

¹Islamic University of Technology, ²York University, ³University of Alberta,

⁴Nanyang Technological University, ⁵Dialpad Inc.

mohsinulkabir@iut-dhaka.edu, saidulis@yorku.ca, tahmid20@yorku.ca,
mnayeem@ualberta.ca, bari0001@e.ntu.edu.sg, enamulh@yorku.ca

Abstract

Large Language Models (LLMs) have emerged as one of the most important breakthroughs in NLP for their impressive skills in language generation and other language-specific tasks. Though LLMs have been evaluated in various tasks, mostly in English, they have not yet undergone thorough evaluation in under-resourced languages such as Bengali (Bangla). To this end, this paper introduces BenLLM-Eval, which consists of a comprehensive evaluation of LLMs to benchmark their performance in the Bengali language that has modest resources. In this regard, we select various important and diverse Bengali NLP tasks, such as text summarization, question answering, paraphrasing, natural language inference, transliteration, text classification, and sentiment analysis for zero-shot evaluation of popular LLMs, namely, GPT-3.5, LLaMA-2-13b-chat, and Claude-2. Our experimental results demonstrate that while in some Bengali NLP tasks, zero-shot LLMs could achieve performance on par, or even better than current SOTA fine-tuned models; in most tasks, their performance is quite poor (with the performance of open-source LLMs like LLaMA-2-13b-chat being significantly bad) in comparison to the current SOTA results. Therefore, it calls for further efforts to develop a better understanding of LLMs in modest-resourced languages like Bengali.

Keywords: Large Language Models, LLM evaluation, Bengali NLP

1. Introduction

Since the advent of pre-trained language models (Devlin et al., 2019; Liu et al., 2019; Rogers et al., 2020), NLP has witnessed revolutionary advancements over the years. These pre-trained language models have produced SOTA results on a variety of NLP tasks with little task-specific fine-tuning. This leads to the development of various pre-trained language models specialized in the Bengali language, such as BanglaBERT (Bhattacharjee et al., 2022), BanglaT5 (Bhattacharjee et al., 2023), and etc. These models have demonstrated exciting progress in many downstream Bengali NLP tasks (Ekram et al., 2022; Akash et al., 2023). However, one major concern for these pre-trained models is that they require fine-tuning using domain-specific large annotated datasets, which is challenging for Bengali due to its under-representation in the NLP domain (Joshi et al., 2020; Chakraborty et al., 2021; Chowdhury et al., 2021) despite being the sixth most spoken language in the world with over 300 million native speakers (Wikipedia, 2023a).

Recent developments in large language models (LLMs) (Brown et al., 2020; Shoeybi et al., 2019; Rae et al., 2022; Zhang et al., 2022) have transformed the landscape in NLP. These LLMs,

with parameter sizes exceeding a hundred billion, leverage the in-context learning capability to achieve impressive performance in few-shot and zero-shot learning scenarios without the need for task-specific fine-tuning. This capability makes it possible to reduce the need for the annotation of huge datasets, allowing the model to perform well on tasks that it was not trained on.

Despite the impressive capabilities of LLMs, they may still frequently generate untruthful facts that diverge from the original input (Ouyang et al., 2022). To address this issue, the Reinforcement Learning from Human Feedback (RLHF) mechanism has been utilized to help LLMs generate honest and harmless responses. ChatGPT and other recently proposed LLMs like PaLM-2, Claude-2, LLaMA-2-chat (Touvron et al., 2023; Anil et al., 2023; Jahan et al., 2024) are trained via leveraging this RLHF technique to mitigate various limitations of the previous generation LLMs and gained widespread popularity. While these LLMs are trained in multiple languages, English possesses the majority of the training data. Though ChatGPT like LLMs has demonstrated strong zero-shot performance in various NLP tasks in English (Laskar et al., 2023a,b) and some other languages (Lai et al., 2023) and domains (Jahan et al., 2023, 2024; Fu et al., 2024), these LLMs are yet to be investigated in the widely spoken, yet modest-resourced, Bengali language domain.

*These authors contributed equally to this work.

†Corresponding Authors.

In this regard, we aim to assess the efficacy of LLMs in prevalent downstream NLP tasks specific to the Bengali language, a domain that has not been thoroughly examined compared to the numerous LLM evaluations conducted for English and other Indo-European languages. Due to the lack of task-specific annotated datasets in Bengali, we conducted a zero-shot evaluation with LLMs to investigate if LLMs could be utilized to address the scarcity of large annotated datasets in Bengali. Thus, the findings from this paper would give insights into the capabilities and limitations of LLMs in Bengali, which may pave the way to implement LLMs in real-world applications like Bengali Chatbots. To this end, we present BenLLM-Eval - a comprehensive benchmark that evaluates the zero-shot performance of various LLMs on diverse NLP tasks in the Bengali language, including text summarization, question answering, paraphrasing, natural language inference, transliteration, text classification, and sentiment analysis. The evaluation incorporates carefully crafted prompts to ensure rigorous assessment of the following three popular LLMs: GPT-3.5, LLaMA-2-13b-chat, and Claude-2, and compare them with SOTA fine-tuned models. To the best of our knowledge, this is the first work that evaluates LLMs in the Bengali language. Our experimental results in this paper are summarized as follows:

- Despite some exceptional cases, the zero-shot performance of LLMs is generally inferior compared to the SOTA fine-tuned models across the majority of the tasks in our evaluation. Given the substantial performance disparities observed, it is reasonable to deduce that LLMs, in their current form, are not suitable for serving as a comprehensive solution for modest-resourced tasks in Bengali¹.
- Considering LLMs remarkable performance in zero-shot scenarios within the English language and its subpar performance in modest-resource languages like Bengali, this paper emphasizes the significance of investigating the limitations of LLMs tailored to diverse modest-resource language groups.

2. Methodology

The objective of our study is to assess the efficacy of LLMs in the context of NLP tasks specific to the Bengali language. We cover 7 diverse and important Bengali NLP tasks: (i) Text Summarization, (ii) Question Answering (QA), (iii) Paraphrasing, (iv)

¹We share the LLM-generated responses, prompts, and parsing scripts for all seven tasks here: <https://github.com/saidul-islam98/BenLLMeval>

Natural Language Inference (NLI), (v) Transliteration, (vi) Text Classification, and (vii) Sentiment Analysis over 8 benchmark datasets. For this purpose, we evaluate ChatGPT (GPT-3.5²), Claude-2³, and LLaMA-2-13b-Chat (Touvron et al., 2023) models. Similar to the prior work on LLM evaluation (Laskar et al., 2023a; Jahan et al., 2023), we also focus on designing a zero-shot learning setting to benchmark the performance of the LLMs.

We prepare a task instruction T for a given test sample X and concatenate the text in the test sample with the task instruction to construct the prompt P . The prompt P is then passed as input to the LLMs, which generates the response R . A comprehensive description of the tasks, datasets, and prompts devised for evaluating each specific task is presented below and also summarized in Table 1. Next, we briefly discuss each task.

Text Summarization: Summarization is the process of automatically generating a concise and coherent summary of a longer text document (Nayeem and Chali, 2017a,b; Laskar et al., 2022), preserving the most important information while reducing the length (Nayeem et al., 2018). In this paper, we evaluate the XL-Sum dataset (Hasan et al., 2021) that consists of 1 million manually annotated data samples from 44 languages. We only took the Bengali samples for evaluation.

Question Answering: For the question-answering task, we evaluate the performance of LLMs on the SQuAD_Bangla dataset (Bhattacharjee et al., 2022). This dataset was constituted using two benchmark English datasets: SQuAD 2.0 (Williams et al., 2018) and TyDi QA (Clark et al., 2020). The objective of this task is to determine whether the answer to a given question Q can be inferred from the reference context C . We provide the reference context along with the question and ask LLMs whether they can infer the answer to the question from the given reference.

Paraphrasing: The paraphrasing task aims to generate a paraphrase of the input text. To evaluate this task, we choose the Bengali samples from the IndicParaphrase dataset (Kumar et al., 2022), which is the largest Indic paraphrasing dataset across 11 Indic languages (Wikipedia, 2023b).

Natural Language Inference: Natural Language Inference (NLI) aims to predict the entailment/contradiction relations between two input sentences: a premise and a hypothesis. To evaluate LLMs for Bengali NLI, we utilize the BNLI dataset (Bhattacharjee et al., 2022) (curated from the benchmark XNLI dataset (Conneau et al., 2018)) that provides annotated data with three categories: *Entailment*, *Contradiction*, and *Neutral*.

Transliteration: Transliteration is the process

²<https://platform.openai.com/docs/models/gpt-3-5>

³<https://www.anthropic.com/index/claude-2>

Dataset	Type	Data Split (Train / Valid / Test)	Prompt
XL-Sum (Hasan et al., 2021)	Text Summarization	8102 / 1012 / 1012	Please provide an one-sentence summary of the following Bangla text input. The input will be a long Bangla paragraph, the output should be a short Bangla paragraph summarizing only the vital information of the input text in one sentence. Please make sure that the output contains the most essential statistical data. Note: Please do not provide anything other than the summarized Bangla output. [INPUT:]
SQuAD Bangla (Bhattacharjee et al., 2022)	Question-Answering	118k / 2.5k / 2.5k	Please provide an answer to the input Bangla question based on the given Bangla context. The input will contain a Bangla question followed by a context. The output should be the answer in Bangla based on the context. Note: Please do not provide anything other than the Bangla answer to the question. [CONTEXT:] [QUESTION:]
IndicParaphrase (Kumar et al., 2022)	Paraphrasing	890k / 10k / 10k	Please provide paraphrasing of the following input Bangla text. The input will be a complex Bangla sentence, the output should be a paraphrased Bangla sentence maintaining the original information of the input text unchanged. Note: Please do not provide anything except the paraphrased Bangla output. [INPUT:]
BNLI (Bhattacharjee et al., 2022)	Natural Language Inference (NLI)	381k / 2.42k / 4.9k	Please determine the logical relationship between the given hypothesis and premise. The input will consist of two sentences written in the Bangla language. The first sentence represents the premise, while the second sentence represents the hypothesis. Your task is to determine whether the hypothesis is false (contradiction), true (entailment), or inconclusive (neutral) given the premise. Please output a number indicating the logical relationship between them: 0 for false (contradiction), 1 for true (entailment), and 2 for inconclusive (neutral) for neutrality. Note: Please avoid providing any additional information beyond the logical relationship. [PREMISE:] [HYPOTHESIS:]
Dakshina (Roark et al., 2020)	Transliteration (single-word: lexicon)	- / - / 9.2k	Task Description: Please provide the transliteration in native Bengali script for the input word. The input will be a word written in Latin script and the output should be the transliterated Bengali word of the given input. Please note that you are not asked to provide translation of the input word, only provide the Bengali transliteration for the given input. Note: Your response should include only the transliterated word in the native Bengali language. Please do not add any explanation with the output. [INPUT:]
Dakshina (Roark et al., 2020)	Transliteration (full sentence)	25k / 5k / 5k	Task Description: Please provide the transliteration in native Bengali script for the input sentence. The input will be a sentence written in Latin script and the output should be the transliterated Bengali sentence of the given input. Please do not provide the translation of the input sentence, only provide the Bengali transliteration for the given input. Note: Your response should include only the transliterated sentence in the native Bengali language. Please do not add any explanation with the output. [INPUT:]
Soham News Article Classification (Kakwani et al., 2020)	Text Classification	11284 / 1411 / 1411	For the Bengali news article given in the input, identify the appropriate section title for the article from the following classes: kolkata, state, sports, national, entertainment, international. Note: Do not output any unnecessary words other than just the section title. The response should be in English language and should be one word. [INPUT:]
IndicSentiment (Doddapaneni et al., 2022)	Sentiment Analysis	- / 156 / 1000	For the given Input, is the sentiment in the input positive or negative? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT:]
SentNoB (Islam et al., 2021)	Sentiment Analysis	12575 / 1567 / 1586	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT:]

Table 1: Datasets Details with our Prompts for Various Tasks.

of converting texts from one script to another based on phonetic similarity. We utilize the *Dakshina dataset* (Roark et al., 2020) to evaluate transliteration errors. Errors can be (i) *Substitutions*, (ii) *Insertions*, or (iii) *Deletions*. A substitution error occurs when the predicted transliteration differs from the reference transliteration. An insertion error happens when extra characters/words are present in the predicted transliteration. A deletion error occurs when characters/words are missing from the predicted transliteration. The error rate is calculated as:

$$\text{error rate} = \frac{\text{substitutions} + \text{deletions} + \text{insertions}}{\text{total reference tokens}} * 100$$

Two types of error rates are proposed for this task in *Dakshina dataset* (Roark et al., 2020): the Character-Error Rate (**CER**) treats each token as an individual Unicode character, while the Word-Error Rate (**WER**) treats each token as a substring separated by whitespace. We report both CER and WER for single-word transliteration, and WER for full-sentence transliteration.

Text Classification: This task refers to the classification of the category for a given input text. To evaluate the text classification capability of LLMs, we use the *Soham Bengali News Classification* dataset that is included in the IndicGLUE (Kakwani

et al., 2020) benchmark. This dataset contains six news categories i.e., *kolkata, state, national, international, sports, and entertainment*.

Sentiment Analysis: We evaluate the Sentiment Analysis capability of the LLMs with two datasets: SentNoB (Islam et al., 2021), and IndicSentiment (Doddapaneni et al., 2022). The SentNoB dataset comprises texts that are informally written in Bengali, collected from public comments on news and videos in social media covering a wide range of domains like, *politics, agriculture, education* etc. The second dataset, IndicSentiment, was manually curated for the IndicXTREME benchmark and contains product reviews of multiple categories. We only used the Bengali samples from the dataset out of the 12 Indic languages.

3. Results and Discussion

We report the performance of LLMs for different tasks and compare their performance with the current SOTA fine-tuned models (see Table 2). Further, we provide an in-depth error analysis for each of the tasks in the Appendix A.

Text Summarization Evaluation: In the XL-Sum dataset, we find that none of the LLMs could outperform the current SOTA mT5 model. Among

Model	XL-Sum (TS)			SQuAD_Bangla (QA)	IndicPara (PP)	BNLI (NLI)	SNAC (TC)	IndicSent (SA)	SentNoB (SA)		
	R-1	R-2	R-L	EM / F1	BLEU	Acc.	Acc.	Acc.	P	R	F1
GPT-3.5	20.19	5.81	15.53	44.85/78.67	2.81	52.71	48.47	90.20	57.70	54.56	53.17
LLaMA-2-13b-chat	0.41	0.14	0.34	31.73/67.95	0.01	42.37	29.27	69.16	48.39	48.49	48.43
Claude-2	20.79	5.55	16.47	49.92/79.04	1.89	32.20	48.61	88.48	53.28	54.38	52.79
mT5 (Hasan et al., 2021)	28.32	11.43	24.23	-	4.45	-	-	-	-	-	-
BanglaBERT (Bhattacharjee et al., 2022)	-	-	-	72.63/79.34	-	82.8	-	-	-	-	-
BanglishBERT (Bhattacharjee et al., 2022)	-	-	-	72.43/78.40	-	80.95	-	-	-	-	-
XLM-R (Large) (Bhattacharjee et al., 2022)	-	-	-	73.15/79.06	-	82.4	-	-	-	-	-
XLM-R (Kakwani et al., 2020; Doddapaneni et al., 2022)	-	-	-	-	-	-	87.60	85.8	-	-	-
IndicBART (Kumar et al., 2022)	-	-	-	-	11.57	-	-	-	-	-	-
IndicBERT (Kakwani et al., 2020; Doddapaneni et al., 2022)	-	-	-	-	-	-	78.45	89.3	-	-	-
mBERT (Kakwani et al., 2020; Doddapaneni et al., 2022)	-	-	-	-	-	-	80.23	72.0	49.58	56.43	52.79
Bi-LSTM + Attn. (w/ FastText) (Islam et al., 2021)	-	-	-	-	-	-	-	-	52.24	63.09	57.15
Bi-LSTM + Attn. (w/ Rand init) (Islam et al., 2021)	-	-	-	-	-	-	-	-	56.16	64.97	60.25

Table 2: Performance Comparison between zero-shot LLMs & fine-tuned SOTA models on Text Summarization (TS), Question Answering (QA), Paraphrasing (PP), Natural Language Inference (NLI), Text Classification (TC), and Sentiment Analysis (SA). EM, Acc., P, R, and F1 denote Exact Match, Accuracy, Precision, Recall, and F1 score respectively. Best results are **boldfaced**.

Task	Pair 6-gram		LSTM		Transformer		Noisy Channel	GPT-3.5		LLaMA-2-13b		Claude 2	
	CER (↓)	WER (↓)	CER (↓)	WER (↓)	CER (↓)	WER (↓)	WER (↓)	CER (↓)	WER (↓)	CER (↓)	WER (↓)	CER (↓)	WER (↓)
Lexicon	14.2	54.0	13.9	54.7	13.2	50.6	-	18.1	60.6	39.85	80.72	23.16	68.07
Sentence	-	39.7	-	-	-	37.6	25.8	-	29.9	-	66.54	-	38.10

Table 3: Single-word and Full-sentence Transliteration results. Here, the baseline results are adopted from Roark et al. (2020). Best results are **boldfaced** and lower (↓) is better.

the LLMs, Claude-2 performs the best in terms of ROUGE-1 and ROUGE-L, while GPT-3.5 performs the best in terms of ROUGE-2. We also find that GPT-3.5 tends to generate much longer summaries (229 words on average) than the gold reference summaries (148 words on average), while the length of the Claude-2 generated summaries (137 words on average) is more consistent with the gold reference summaries (148 words on average). Moreover, to find out why LLaMA-2-13b-chat achieves very low ROUGE scores, we manually reviewed the responses and found out that it ended up generating all the summaries in English. By translating the summaries to Bengali using Google’s Translation API (Google, 2020), the R-1, 2, and L scores are increased to 4.69, 0.61, and 3.61, respectively, for LLaMA-2-13b-chat.

Question Answering Evaluation: Since LLMs frequently generate responses that may not be an exact match of the gold label but are nonetheless correct, our QA evaluation utilizes human intervention to compute the F1 score. We find that Claude-2 and GPT-3.5 achieve performance almost similar to the current SOTA results in terms of F1. However, the performance of all LLMs is quite poor in terms of Exact Match due to generating descriptive responses, as well as paraphrases.

Paraphrasing Evaluation: For the paraphrasing task, we observe a very low BLEU score for all LLMs, which is a phenomenon similar to what happened with the EM metric for the QA task. Note that the BLEU score also computes word-level similarity, while such limitations of word-based similarity metrics have also been noticed LLMs are evaluated in English datasets (Laskar et al., 2023a).

Natural Language Inference Evaluation: We observe that while GPT-3.5 achieves the best results among all LLMs on the BNLI dataset, it is still much lower in comparison to the current SOTA

BanglaBERT. Surprisingly, we find that Claude-2 is the worst performer in this task while being outperformed by the much smaller LLaMA-2-13b-chat model. To further investigate in which cases LLMs perform poorly in the NLI task, we demonstrate their predictions using the Confusion Matrix in Figure 1 of Appendix and find that LLaMA-2-13b-chat was quite bad while predicting the *Neutral* types and Claude-2 performed poorly while predicting the *Entailment* types. In the case of GPT-3.5, it performed poorer while predicting the *Neutral* type class in comparison to the other types.

Transliteration Evaluation: From Table 3, we find that GPT-3.5 achieves superior performance over other LLMs in single-word transliteration. While it fails to outperform the existing results, it still achieves competitive performance compared to the prior SOTA models. This is an interesting observation given the intricate structure and infrequent usage of lexicons in the test set. In full-sentence transliteration, GPT-3.5 makes fewer word errors and character errors (60.6% & 18.1%) compared to other LLMs ((80.72% & 39.85%) for Llama-2-13b-chat and (68.07% & 23.16%) for Claude-2). However, the results are quite poor compared to other SOTA models.

Text Classification Evaluation: LLMs performed poorly on the Soham News Article classification dataset, with the best-performing LLM in this dataset, the Claude-2 model achieving only 48.61% accuracy, followed by GPT-3.5 achieving 48.47% accuracy while the LLaMA-2-13b-chat being the worst performer with an accuracy of 29.27%. On the contrary, the XLM-R model, which is the SOTA for this task, obtained an accuracy of 87.60% in this dataset. While we achieved these results by providing the class labels alongside our prompts (see Table 1), if the class labels were not mentioned, the accuracy dropped to 20.76%,

18.36%, and 1.06% for Claude-2, GPT-3.5, and LLaMA-2-13b-chat, respectively. This indicates the importance of using descriptive prompts to improve the zero-shot performance of LLMs.

Sentiment Analysis Evaluation: In the sentiment classification task, we observe that GPT-3.5 performed exceptionally well on the IndicSentiment dataset (Doddapaneni et al., 2022), attaining a new SOTA accuracy of 90.20%. While we also find that the Claude-2 model achieved impressive performance (88.48% accuracy), the LLaMA-2-13b-chat model performed much poorer. We also find a similar trend for sentiment analysis in the SentNoB (Islam et al., 2021) dataset. In this dataset, to ensure a fair comparison with the current SOTA, we use the precision, recall, and F1 scores as evaluation metrics. While none of the LLMs could outperform the current SOTA, all of them (especially GPT-3.5 and Claude-2) achieved performance comparable to the SOTA results. This is consistent with LLMs' performance for this task in English (Laskar et al., 2023a).

A detailed error analysis and sample outputs for all tasks are presented in Appendix A.

4. Task Contamination Analysis

Task contamination analysis (Li and Flanigan, 2023) is essential to ensure a fair model evaluation since it helps identify a model's prior exposure to test tasks on its training data. Inspired by the work of Li and Flanigan (2023), we include task contamination analysis in our evaluation to appropriately assess the performance of the LLMs. We utilize two methods: *Task Example Extraction (TEE)* and *Membership Inference (for generative tasks like summarization and paraphrasing)* to verify the evidence of task contamination. **TEE** involves retrieving task examples from instruction-tuned models, although it doesn't require an exact match with existing training data, helps identify potential task contamination in zero and few-shot learning scenarios, as any task examples found indicate possible prior exposure. Additionally, **Membership Inference** is a method used for generation tasks, where the model's generated content is checked against the original dataset for exact matches. If a match is found, it implies that the content was part of the LLM's training data, indicating direct exposure rather than general learning ability.

At first, we applied TEE to all tasks without explicitly mentioning the dataset name. Our findings reveal that only GPT-3.5 could generate examples related to these tasks (except Natural Language Inference), while Claude-2 and LLaMA-2-13b-chat models failed to extract task examples for any tasks. Therefore, there is a possibility that such tasks were already included in the pre-

training data of GPT-3.5. Regarding the BNLI dataset where no models could extract any task examples, we find that the premises, hypotheses, and labels generated by all LLMs for Bengali were significantly inaccurate, providing evidence that contamination did not occur. In terms of extracting task examples in the transliteration task, we find that only GPT-3.5 could extract the task examples for both word-level and sentence-level transliteration, whereas both LLaMA-2-13b-chat and Claude-2 failed to extract any task examples.

Finally, we applied the Membership Inference technique in the generative tasks (e.g., summarization and paraphrase generation) to analyze task contamination. In the summarization task, none of the LLMs produced an output that aligned with the test data labels. However, on the paraphrasing task, GPT-3.5 produced around 50 exact match instances, while Claude-2 produced 30 and LLaMA-2-13b-chat produced 15 exact matches of the generated outputs and test labels. So it is possible that the test set of the IndicPara dataset was exposed to the LLMs while training.

In summary, contamination could be an issue with the GPT-3.5 model in Sentiment Analysis, Text Classification, Summarization, and QA tasks, while all the models, i.e., GPT-3.5, LLaMA-2-13b-chat, and Claude-2 were affected by task contamination in the Paraphrasing task. However, in Natural Language Inference, we did not see any evidence of task contamination.

5. Conclusions and Future Work

In this paper, we introduce BenLLM-Eval, which provides a comprehensive zero-shot evaluation of LLMs on seven benchmark NLP tasks to understand the capability of LLMs in the modest-resourced Bengali language. The results revealed that while in some tasks, zero-shot closed-source LLMs like GPT-3.5 or Claude-2 perform on par (e.g., summarization) or even outperform (e.g., sentiment analysis) current SOTA models, in most tasks the LLMs exhibited considerably lower performance compared to supervised fine-tuned models. We also observed that the open-source LLaMA-2-13b-chat model performed significantly poorer in most tasks. Thus, open-source LLMs should be extensively evaluated on low to modest-resource languages to ensure a proper understanding of their capabilities and limitations. Our findings also reveal that in order to achieve optimal performance across languages, LLMs should also be trained on corpus covering various languages along with English. In the future, we intend to expand our experiments by including additional low to modest-resource languages, tasks, datasets, and settings to obtain broader insights.

Limitations

The datasets used to train LLMs like GPT-3.5 or Claude-2 are not disclosed; consequently, some datasets used to evaluate the model may or may not have existed during model training. In addition, the API used to evaluate the GPT-3.5 model is based on OpenAI's *GPT-3.5 turbo*. Although an improved version GPT-4 is also available, GPT-4 is significantly more expensive than GPT-3.5. Thus, we could not use GPT-4 in this work. Other than these limitations, this study will provide a practical direction for future research into the performance of LLMs in modest-resourced languages in the NLP domain.

Ethics Statement

This study assesses the performance of LLMs in seven benchmark NLP tasks in the Bengali language wherein LLMs are tasked with generating an output based on the information given in the input. Therefore, no prompt or information was presented to LLMs that could result in the generation of any output raising any ethical or privacy issues. Furthermore, since this study is targeted only to the evaluation of the effectiveness and limitations of LLMs in Bengali NLP downstream tasks, conducting the study does not lead to any unwanted biases. There will be no licensing issue as all the datasets that are utilized in this research are publicly available.

Acknowledgements

This research is supported by the Generic (Minor/Startup/Other) research funds of York University awarded to Md Tahmid Rahman Laskar. We also thank Anthropic for providing free access to the Claude-2 model, the Canada Foundation for Innovation grant and Compute Canada for its computing resources. Mir Tafseer Nayeem is also supported by the Huawei PhD Fellowship.

References

Abu Ubaida Akash, Mir Tafseer Nayeem, Faisal Tareque Shohan, and Tanvir Islam. 2023. [Shironaam: Bengali news headline generation using auxiliary information](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 52–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey,

Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubashir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. [BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. [BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Susmoy Chakraborty, Mir Tafseer Nayeem, and Wasi Uddin Ahmad. 2021. [Simple or complex? learning to predict readability of bengali texts](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12621–12629.

Radia Rayan Chowdhury, Mir Tafseer Nayeem, Tahsin Tasnim Mim, Md. Saifur Rahman Chowdhury, and Taufiqul Jannat. 2021. [Unsupervised abstractive summarization of Bengali text documents](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2612–2619, Online. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. [TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages](#).

- Transactions of the Association for Computational Linguistics*, 8:454–470.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Sumanth Doddapaneni, Rahul Aralikkatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. [Indicxtreme: A multi-task benchmark for evaluating indic languages](#).
- Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. [BanglaRQA: A benchmark dataset for under-resourced Bangla language reading comprehension-based question answering with diverse question-answer types](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2518–2532, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Bhushan TN. 2024. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? *arXiv preprint arXiv:2402.00841*.
- Google. 2020. [Googletrans: Free and unlimited google translate api for python](#). Accessed: March 12, 2024.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. [XL-sum: Large-scale multilingual abstractive summarization for 44 languages](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.
- Khondoker Ittehadul Islam, Sudipta Kar, Md Saiful Islam, and Mohammad Ruhul Amin. 2021. [SentNoB: A dataset for analysing sentiment on noisy Bangla texts](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3265–3271, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. [Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers](#). In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.
- Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, page 108189.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Aman Kumar, Himani Shrotriya, Prachi Rani Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. 2022. [Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. [Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning](#).

- Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023a. [A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023b. [Building real-world meeting summarization systems using large language models: A practical perspective](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics.
- Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.
- Changmao Li and Jeffrey Flanigan. 2023. [Task contamination: Language models may not be few-shot anymore](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pre-training approach. *ArXiv*, abs/1907.11692.
- Mir Tafseer Nayeem and Yllias Chali. 2017a. [Extract with order for coherent multi-document summarization](#). In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 51–56, Vancouver, Canada. Association for Computational Linguistics.
- Mir Tafseer Nayeem and Yllias Chali. 2017b. [Paraphrastic fusion for abstractive multi-sentence compression generation](#). In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17*, page 2223–2226, New York, NY, USA. Association for Computing Machinery.
- Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. [Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. [Is chatgpt a general-purpose natural language processing task solver?](#)
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. [Scaling language models: Methods, analysis & insights from training gopher](#).
- Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. [Processing South Asian languages written in the Latin script: the Dakshina dataset](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. [A primer in BERTology](#).

What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. [Megatron-lm: Training multi-billion parameter language models using model parallelism](#). *CoRR*, abs/1909.08053.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutli Bhosale, et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *arXiv preprint arXiv:2307.09288*.

Wikipedia. 2023a. [Bengali language](#). Accessed: May 23, 2023.

Wikipedia. 2023b. [Indic languages](#). Accessed: May 23, 2023.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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. 2022. [Opt: Open pre-trained transformer language models](#).

A. Appendix

A.1. Prompt Construction Details

In the process of adapting datasets into prompts, we carefully analyzed the prompt templates in various LLM evaluation papers on English datasets (Laskar et al., 2023a; Qin et al., 2023) and then explored several prompt variants in Bengali for each task to ascertain the optimal one. For example, at the beginning of the Soham News Article Classification task, our prompts did not feature class labels, which caused a pronounced failure in the accurate classification of news articles by all tested LLMs, markedly degrading their effectiveness. However, this situation improved considerably when we updated the prompts to integrate class labels, leading to a notable improvement in

the models' capacity to correctly associate articles with their respective classes. Thus, experimenting with different variations of prompts to come up with the optimal one is a crucial step in our prompt design phase.

A.2. In-depth Error Analysis

In this section, we discuss some interesting cases where GPT-3.5 failed to generate coherent responses given the input prompt. We discuss these cases in the following sections:

A.3. Generative Tasks

The error cases for generative tasks like abstractive summarization, question answering, and paraphrasing is discussed in this section.

A.4. Abstractive Summarization

In the summarization task, we notice that GPT-3.5's responses are highly inconsistent. While GPT-3.5 occasionally recognizes the context and provides a good summary of the key points, it frequently misses these details in the reference text and overstuffs the response. One example can be demonstrated as follows:

Prompt: Please provide an one-sentence summary of the following Bangla text input. The input will be a long Bangla paragraph, the output should be a short Bangla paragraph summarizing only the vital information of the input text in one sentence. Please make sure that the output contains the most essential statistical data. Note: Please do not provide anything other than the summarized Bangla output.

Input: মানসুর আব্বাস প্রধানমন্ত্রী বা তার প্রতিপক্ষ - কেউই আসলে ক্ষমতায় যাবার মতো নিরঙ্কুশ সংখ্যাগরিষ্ঠতা পাননি। আর এর মধ্যেই কিংমেকার হয়ে উঠেছে রাম (Raam) নামে একটি ইসলামপন্থী আরব দল (ইউনাইটেড আরব লিস্ট হিসেবেও দলটি পরিচিত)। এবারের নির্বাচনে পাঁচটি আসনে জয় পেয়েছে এ দলটি- যা মিস্টার নেতানিয়াহকে ক্ষমতায় রাখা বা ক্ষমতা থেকে বিদায় করে দিতে ভূমিকা রাখার সুযোগ করে দিয়েছে। বিবিসি বাংলায় আরও পড়ুন: ফিলিস্তিনি-ইসরায়েলি সংঘাতের মূলে যে দশটি প্রশ্ন ইসরায়েলকে ইহুদি রাষ্ট্র ঘোষণার প্রতিক্রিয়া কী হবে? ইসরায়েলের কাছে আরবরা কেন পরাজিত হয়েছিল? ২০১৯ সাল থেকে নেতৃত্ব দিচ্ছেন মানসুর আব্বাস রাম: বিস্ময়কর কিংমেকার রক্ষণশীল মুসলিম মানসুর আব্বাসের নেতৃত্বে রাম দলটি মূলত ফিলিস্তিনের গাজা শাসন করা হামাসের ধর্মীয় ভাবধারায় গড়ে ওঠা একটি ইসলামপন্থী দল। ১৯৯৬ সালে প্রতিষ্ঠিত এই দলটি শুরু থেকে ইসরায়েলের পার্লামেন্ট আসন পেয়ে আসছে। যদিও ২০০৯ সালের নির্বাচনে অংশ নেয়ার ক্ষেত্রে তাদের ওপর নিষেধাজ্ঞা দিয়েছিলো দেশটির নির্বাচনী কর্তৃপক্ষ। পরে সুপ্রিম কোর্ট ওই নিষেধাজ্ঞা বাতিল করে দেয়। ২০২০ সালে ইসরায়েলের আরব রাজনৈতিক দলগুলোর জোট- জয়েন্ট আরব লিস্টের অংশ ছিলো যারা নজিরবিহীন ভাবে পনেরটি সিট পেয়েছে পার্লামেন্টে। তবে একাই নির্বাচনে লড়ার জন্য চলতি বছরের ২৮শে জানুয়ারি ওই জোট ছেড়ে আসে রাম। আরব সম্প্রদায়ের সামগ্রিক দরকষাকষির

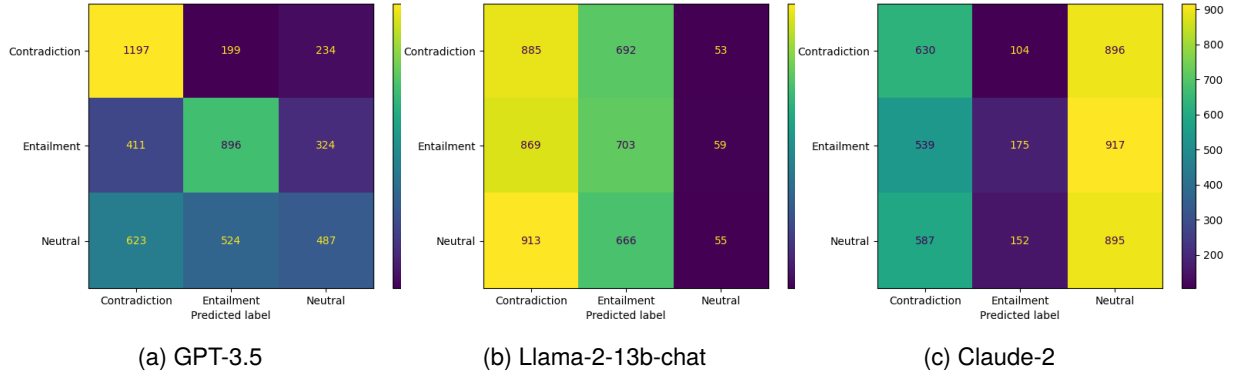


Figure 1: Confusion matrices for different LLMs on the BNLI dataset for the NLI task.

ক্ষমতাকে হুমকির মুখে ফেললেও এ সিদ্ধান্ত দলটির 'কিংমেকার' হয়ে ওঠার পথও তৈরি করে দিয়েছে। নির্বাচনে ভোট দিচ্ছেন এক ইসরায়েলি-আরব নারী দলটির পেছনে থাকা ব্যক্তি ৪৬ বছর বয়সী মানসুর আব্বাস হলেন দলটির মূল ব্যক্তি। ইসরায়েলের রাজনৈতিক ইতিহাসের গুরুত্বপূর্ণ সময়টিতে তিনি এখন রাজনীতির কেন্দ্রে রয়েছেন। জেরুজালেমের হিব্রু বিশ্ববিদ্যালয় থেকে পাশ করা দস্ত চিকিৎসক তিনি। যদিও পরে হাইফা বিশ্ববিদ্যালয় থেকে রাষ্ট্রবিজ্ঞানে মাস্টার্স ডিগ্রি লাভ করেন। ২০১৯ সালে তিনি ইউনাইটেড আরব লিস্টের নেতা মনোনীত হন এবং ওই জোটের অংশ হিসেবে পার্লামেন্টেও নির্বাচিত হন। অবশ্য বেনিয়ামিন নেতানিয়াহু ও তার লিকুদ পার্টির সাথে যোগাযোগ বাড়ানোর চেষ্টা করতে গিয়ে তিনি দলের মধ্যে বিভক্তিও তৈরি করেন। ইসরায়েলে আরব রাজনীতি ইসরায়েলের প্রায় নব্বই লাখ মানুষের মধ্যে উনিশ লাখের মতো আরব আছেন যারা ১৯৪৮ সালে ইসরায়েলের সীমানায় থেকে গিয়েছিলেন। ওই প্রায় সাড়ে সাত লাখ ফিলিস্তিনি তাদের বাড়িঘর ছেড়ে পালিয়েছিলো অথবা তাদের বহিষ্কার করা হয়েছিলো। যদিও ইসরায়েলের নাগরিক এমন অনেক আরব নিজেদের ফিলিস্তিনি বা ইসরায়েলি ফিলিস্তিনি হিসেবে পরিচিত হতে পছন্দ করেন আর অন্যরা নিজেদের ইসরায়েলি আরব হিসেবে উল্লেখ করে থাকে। ইসরায়েলের আরবদের বেশিরভাগই সুন্নি মতাদর্শে বিশ্বাসী আর দ্বিতীয় বড় অংশটি খ্রিস্টান। ইসরায়েলের দশ শতাংশ মুসলিম আরব বেদুইন গোত্র থেকে আসা। ১৯৪৯ সালের নির্বাচনে তারা অংশ নিয়েছিলো। আরব রাজনৈতিক দলগুলো দেশটিতে আরবদের সমান অধিকারের পক্ষে সোচ্চার ভূমিকা পালন করে এবং একই সাথে ফিলিস্তিনের প্রতিও তাদের সমর্থন আছে। বেনজামিন নেতানিয়াহু এখন রাম কোনো জোটে যাবে? গত বছর মিস্টার নেতানিয়াহুর প্রতিপক্ষ বেনি গান্টজ আরব দলগুলোর সাথে জোট করে সরকার গঠনের বিষয়টি প্রত্যাখ্যান করেছিলেন। এর আগে নেতানিয়াহু নিজেও রামের সাথে জোটের সম্ভাবনা উড়িয়ে দিয়েছিলেন। যদিও প্রচারণায় তার সুর ছিলো নরম। এখন যদি তারা একটি জোট করতে পারেন তাহলে মিস্টার নেতানিয়াহু ও মিস্টার আব্বাস- দুজনেই লাভবান হবেন। যদিও সংবাদদাতারা বলছেন, নেতানিয়াহুর অন্য শরীকদের সাথে তারা কিভাবে একযোগে কাজ করবেন তা এখনো পরিষ্কার নয়। জেরুজালেমে বিবিসি সংবাদদাতা অবশ্য বলছেন কোয়ালিশনে না থেকেও মিস্টার নেতানিয়াহুকে সমর্থন দেয়ার চুক্তি করতে পারে রাম। কিন্তু চুক্তি বা সমঝোতা যাই হোক- মিস্টার আব্বাস এখন নিশ্চিত যে দরকষাকষির সুযোগ তার হাতে। তিনি বলছেন ইসরায়েলের আরব জনগোষ্ঠীর জন্য সেরা সিদ্ধান্তটিই তিনি নেবেন। যদিও অনেকেই এখনো এ নিয়ে নিশ্চিত নন। জেরুজালেম অধিকার কর্মী ইতাফ আওয়াদ বলছেন, লিকুদ ও রাম দলের জোট ফিলিস্তিনীদের জন্য ভালো হবে না।

Expected Response: গত দু বছরের মধ্যে চতুর্থবারের

মতো একই সমস্যা হচ্ছে ইসরায়েলে- নির্বাচনের পর সংখ্যাগরিষ্ঠতা পেয়েও ক্ষমতায় থাকার জন্য প্রয়োজনীয় আসনের জন্য হাফাকার করতে হচ্ছে প্রধানমন্ত্রী বেনিয়ামিন নেতানিয়াহুকে।

GPT-3.5 Response: মানসুর আব্বাস নেতৃত্বের রাম দল ইসরায়েলের পার্লামেন্টে পাঁচটি আসন জিতেছে এবং এদের কিংমেকার হয়ে উঠার পথ তৈরি করেছে, তিনি ইউনাইটেড আরব লিস্টের নেতা হিসেবে পরিচিত এবং ইসরায়েলে আরব রাজনীতি প্রায় নব্বই লাখ মানুষের মধ্যে উনিশ লাখের মতো আরব আছেন যারা ১৯৪৮ সালে ইসরায়েলের সীমানায় থেকে গিয়েছিলেন।

In the aforementioned example, GPT-3.5 fails to highlight the most crucial information from the input and saturates the output with superfluous texts.

A.4.1. Question Answering

In the Question Answering task, GPT-3.5 is given instructions to provide responses based on the given context. However, there are instances where GPT-3.5 deviates from the context and attempts to answer using information from its general knowledge, resulting in a mismatch with the gold label. While this problem occurred only in a limited number of cases, the authors were unable to independently verify the accuracy of the responses in these instances, leading to their classification as incorrect answers.

Prompt: Please provide an answer to the input Bangla question based on the given Bangla context. The input will contain a Bangla question followed by a context. The output should be the answer in Bangla based on the context. Note: Please do not provide anything other than the Bangla answer to the question.

Context: নেপালের রাজধানী কাঠমুন্ডুর উপত্যকায় প্রাপ্ত নিওলিথিক যুগের বেশকিছু উপাদান এটিই নির্দেশ করে যে হিমালয়ান অঞ্চলে প্রায় ৯০০০ বছর থেকে মানুষ বসবাস করছে। এটি প্রতিষ্ঠিত যে প্রায় ২৫০০ বছর পূর্বে নেপালে তিব্বতী-বার্মীয় জনগোষ্ঠীর বসবাস ছিল। ১৫০০ খৃস্টপূর্বাব্দে ইন্দো ইরানীয় বা আর্য জাতিগোষ্ঠী এই হিমালয়ান উপত্যকায় প্রবেশ করে। ১০০০ খৃস্টপূর্বাব্দের দিকে এই অঞ্চলটিতে বিভিন্ন গোষ্ঠীর জন্য স্বতন্ত্র রাজ্য ও কনফেডারেশন গড়ে উঠে। এরকমই একটি কনফেডারেশন ছিল সাকিয়া যার একসময়কার রাজা ছিলেন সিদ্ধার্থ গৌতম (৫৬৩-৪৮৩ খৃস্টা-

পূর্বাব্দ) যিনি গৌতম বুদ্ধ বা শুধু বুদ্ধ নামেই পরিচিত। তিনি পবিত্র ও সাধনাময় জীবনযাপনের জন্য তার রাজত্ব ত্যাগ করেছিলেন। ২৫০ খ্রিস্টপূর্বাব্দে এই অঞ্চলটি উত্তর ভারতের মৌর্য সাম্রাজ্যের (Mauryan) অধীনে আসে এবং পরবর্তীতে ৪র্থ শতাব্দীতে এটি গুপ্ত সাম্রাজ্যের অধীনে একটি পুতুল রাষ্ট্রে পরিণত হয়। পঞ্চম শতাব্দীর শেষ হতে শুরু করে পরবর্তী বেশ কিছুটা সময় শাসন করে একদল শাসক যারা সাধারণভাবে লিচ্ছবি (Licchavis) নামে পরিচিত। লিচ্ছবি সাম্রাজ্যের (Licchavi) পতন ঘটে অষ্টম শতাব্দীতে এবং এরই সাথে শুরু হয় নেওয়ারি (Newari) যুগের। ৮৭৯ সালে নেওয়ারিদের রাজত্ব শুরু হলেও সমগ্র রাষ্ট্রের উপর তাদের নিয়ন্ত্রণ অনেকটাই অনিশ্চিত ছিল। একাদশ শতাব্দীর শেষ ভাগে নেপালের দক্ষিণাংশ দক্ষিণ ভারতের চালুক্য সাম্রাজ্যের (Chalukaya) অধীনে আসে। চালুক্যদের রাজত্বকালে নেপালের ধর্মে ব্যাপক পরিবর্তন আসে কারণ সব রাজাই হিন্দু ধর্মের পৃষ্ঠপোষকতা করতেন এবং বৌদ্ধ ধর্মের প্রসারের বিপরীতে হিন্দু ধর্মের প্রচারে অবদান রাখেন।

Question: নেপাল কি একটি সার্বভৌম দেশ ?

Expected Response: No Answer.

GPT-3.5 Response: হ্যাঁ, নেপাল একটি সার্বভৌম দেশ। In this particular instance, GPT-3.5 provides responses that go beyond the scope of the given context, as the context itself lacks the necessary information to yield the correct answer. Consequently, this deviation from the context leads to a mismatch with the gold label, ultimately resulting in an incorrect response.

A.4.2. Paraphrasing

In some instances, GPT-3.5 generates paraphrase that completely changes the meaning of the input sentences. For example:

Input: আগুনে কোনও হতাহতের খবর পাওয়া যায়নি। [No injury was reported because of the fire]

Expected Responses: ['অগ্নিকাণ্ডে উল্লেখযোগ্য কোন ক্ষয়ক্ষতি হয়নি।', 'অগ্নিকাণ্ডে উল্লেখযোগ্য কোন ক্ষয়ক্ষতি হয়নি।', 'এই অগ্নিকাণ্ডে হতাহতের কোনও খবর মেলেনি।', 'এই অগ্নিকাণ্ডে হতাহতের কোনো খবর পাওয়া যায়নি।', 'অগ্নিকাণ্ডে হতাহতের কোনো খবর পাওয়া যায়নি।', 'অগ্নিকাণ্ডে কারও হতাহতের খবর পাওয়া যায়নি।']

GPT-3.5 Response: কোন আপত্তিকর ঘটনা না হওয়ায় আগুনের সম্পর্কে কোন খবর পাওয়া যায়নি। [There was no news about the fire as there was no untoward incident.] This is a matter of concern and presents additional evidence that GPT-3.5 is not currently suitable for serving as a universal problem solver in the Bangla language.

A.5. Classification Tasks

In this section, we discuss the error cases in the other three tasks that we consider: Natural Language Inference (NLI), Text Classification, and Sentiment Analysis.

A.5.1. Natural Language Inference

The confusion matrix obtained by evaluating GPT-3.5 for the NLI task is demonstrated in Figure

1. The matrix reveals that GPT-3.5 demonstrates high accuracy in predicting the *Contradiction* and *Entailment* labels, but encounters significant challenges in accurately predicting the *Neutral* labels. Approximately 49% of the misclassifications arise when attempting to predict the *Neutral* class. A thorough manual examination uncovers that GPT-3.5 often exhibits bias towards expressing a particular opinion polarity (*Contradiction*, *Entailment*) when dealing with logical relationships in Bangla, and it fails to appropriately recognize and convey neutrality even in cases where it is evident.

Prompt: Please determine the logical relationship between the given hypothesis and premise. The input will consist of two sentences written in the Bangla language. The first sentence represents the premise, while the second sentence represents the hypothesis. Your task is to determine whether the hypothesis is false (contradiction), true (entailment), or inconclusive (neutral) given the premise. Please output a number indicating the logical relationship between them: 0 for false (contradiction), true (entailment), and 2 for inconclusive (neutral) for neutrality. Note: Please avoid providing any additional information beyond the logical relationship.

Premise: ঠিক আছে, যেখানে দুই-তিনটা বিমান এক সপ্তাহে এসে পৌঁছেছে এবং আমি জানতাম না তারা কোথায় যাচ্ছে।

Hypothesis: প্রতি সপ্তাহে একটিরও বেশি বিমান আসে।

Expected Response: 1 (Entailment)

GPT-3.5 Response: 2 (Neutral)

A.5.2. Transliteration

A deeper look into the cases where LLMs fail yields the following observations:

(i) **Intricate pronunciation rules in Bengali pose challenges in transliteration.** Bengali, being a language rich in nuances, possesses distinct rules for pronunciation and spelling. In Bengali, the same letter may produce different sounds when used in different words. All LLMs encounter significant challenges in scenarios where specific rules govern the pronunciation of a particular letter. For instance, the Latin lexicon 'andre' is pronounced as 'আন্দ্রে' in Bengali (similar to 'addre' in English). The 'n' letter of the lexicon 'andre' is silent in Bengali and instead represented as 'ঁ', which creates a nasal sound, whereas even the response of GPT-3.5 (which is the best-performing LLM in this task) is 'আন্দ্রে', which includes an 'n' sound. Such unconventional cases represent the primary reasons for LLMs to struggle while transliterating lexicons from Latin to Bengali.

(ii) **Difficulty in transliterating lexicons that closely resemble English words.** For instance, when encountering the term 'cador', which bears a strong resemblance to the English word 'cadre', GPT-3.5 mistakenly transliterates it as 'ক্যাডর', a

transliteration that shares a similar pronunciation to 'cadre'. However, the correct gold label for this particular example is 'চাদর'. Such instances of erroneous substitution lead to an increase in CER.

(iii) Transliteration of outdated pronunciations. Languages continuously undergo changes based on their usage by individuals. In the Dakshina dataset, certain gold labels are annotated with outdated pronunciations. While LLMs correctly transliterate various lexicons into more contemporary pronunciations, many such instances in the Dakshina dataset are considered incorrect transliterations due to outdated gold labels. For instance, the term 'ssoshur' is transliterated as 'শুশুর' by GPT-3.5, which represents an accurate contemporary transliteration. Nevertheless, the gold label for this example is given as 'শুশুর'. Such cases result in 90% of the substitution errors in GPT-3.5.

Full Sentence Transliteration: As seen from table 3, the zero-shot GPT-3.5 on full sentence transliteration surpasses all the SOTA models except one (i.e., *Noisy Channel*). Meanwhile, Claude-2 achieves comparable performance with the prior models. The error patterns encountered in full-sentence transliteration largely align with those observed in single-word transliteration, with the addition of the following error case:

(i) Code-mixing creates a stumbling block for GPT-3.5. While transitioning between vocabulary words, LLMs demonstrate a tendency to make mistakes. For instance, the following sentence 'Amake to tomra *mute* kore rekhechile' contains an occurrence of code-mixing with the English word 'mute'. GPT-3.5 misidentifies it as a Bengali word and transliterates the sentence as 'আমাকে তো তোমরা মুটে করে রেখেছিলে', whereas the accurate transliteration of the word 'mute' should be 'মিউট' instead of 'মুটে'. Such cases contribute to substitution errors in full-sentence transliteration.

A.5.3. News Article Classification

Our experimental results show that GPT-3.5 misclassified the category *kolkata* the most, failing to generate the correct response in 503 of the test set's 569 examples (88.4%). The subsequent category with the highest frequency of misclassification is *national*. GPT-3.5 was unable to accurately classify this particular category in 95 out of 175 instances, representing a misclassification rate of 54.20%. On the other hand, GPT-3.5 effectively identified the category labeled as *entertainment* in 110 out of 130 occurrences, resulting in a success rate of 92.30%.

A.5.4. Sentiment Analysis

The examples listed in Table 5, illustrate the error cases where GPT-3.5 misclassifies the sentiment

of the given input. The first seven examples are from the IndicSentiment test set and the rest of the examples are from the SentNoB dataset. Notably, the *positive* class exhibits the most frequent misclassification in both test sets, suggesting that GPT-3.5 still has a considerable way to go toward attaining a complete understanding of the Bangla language. In the instances where the class is *neutral*, the examples are mostly simple statements with no sentiment word associated with them (table entries 11, and 12), yet GPT-3.5 classified them as either *negative* or *positive*. Furthermore, GPT-3.5 demonstrates difficulties in capturing challenging Bangla sentiment words, i.e., অমায়িক (*Accommodating* in English), সুবিধা (*Advantage* in English), আপসহীন (*Uncompromising* in English) etc (example 2, 8, 13 in Table 5).

A.6. Unexpected, out of range response

In the tasks of text classification and sentiment analysis, the prompts are designed to include the target classes in order to identify any unexpected or non-class responses. Based on our experimental findings, we observe that the outputs generated by GPT-3.5 exhibit outcomes that deviate from the expected range of outputs.

News Article Classification.

Prompt: For the Bengali news article given in the input, identify the appropriate section title for the article from the following classes: kolkata, state, sports, national, entertainment, international. Note: Do not output any unnecessary words other than just the section title. The response should be in English language and should be one word. Input : ১০০ দিনের কাজের প্রকল্পের সাফল্য দেখতে রাজ্যে আসছেন কেন্দ্রের প্রতিনিধি। সঙ্গে ১০টি রাজ্যের প্রতিনিধিও। ৯ থেকে ১২ জানুয়ারির মধ্যে পাঁচটি জেলায় রাজ্যের 'উদ্ভাবনী কাজের' সাফল্য দেখবেন তাঁরা। ঘটনাচক্রে, পঞ্চায়েত নির্বাচনের আগে এই সফর হতে চলেছে। ফলে প্রশাসনের একাংশের মতে, এই সফর ভোটের আগে শাসকদলের কাছে বড় প্রাপ্তি হতে পারে। বিশেষত, রাজ্যগুলির মধ্যে অধিকাংশই যখন বিজেপি অথবা বিজেপি'র সহযোগী দলের সরকার পরিচালিত। 'ইন্টার স্টেট এক্সচেঞ্জ প্রোগ্রামের' আওতায় প্রতি বছর একটি বা দুটি রাজ্যকে বেছে নেওয়া হয়। চলতি বছরে এ রাজ্যকে বেছে নিয়েছে কেন্দ্র। কেন্দ্রীয় সরকারের যুগ্মসচিব (এনআরইজিএ) অপরাধিতা সারেন্দির নেতৃত্বে ১১ জন পদস্থ আধিকারিক আসবেন। থাকবেন ১০টি রাজ্য থেকে পাঁচজন করে সরকারি প্রতিনিধি। রাজ্যগুলির মধ্যে রয়েছে তামিলনাড়ু, অন্ধ্র, তেলঙ্গনা, কর্ণাটক, ওড়িশা, বিহার, হিমাচলপ্রদেশ, ছত্তীসগড় এবং মধ্যপ্রদেশ। জম্মু ও কাশ্মীর প্রতিনিধি পাঠাবে কি না, তার নিশ্চয়তা মেলেনি। পাঁচটি দলে ভাগ করে প্রতিনিধিদের নিয়ে যাওয়া হবে পশ্চিম বর্ধমান, বাঁকুড়া, পুরুলিয়া, ঝাড়গাম এবং বীরভূমে। ১০০ দিনের কাজের প্রকল্পের আওতায় শুকিয়ে যাওয়া ঝরনায় বিকল্প উপায়ে জলের জোগান ফিরিয়ে আনা, ফ্লাই অ্যাশের ইট তৈরি, পুরুলিয়ার মাটিতে বেদানার ফলন, মাটির চরিত্র ভিন্ন হলেও আরবের খেজুর চাষ, লাতিন আমেরিকা এবং দক্ষিণ পূর্ব এশিয়ার জনপ্রিয় ড্রাগন ফুটের চাষ, বৃক্ষপাট্টা দিয়ে গাছ বাঁচানো-সহ বিভিন্ন কাজ প্রতিনিধিদলের সামনে তুলে

ধরা হবে। এই অভিজ্ঞতা ভবিষ্যতে কীভাবে ছড়িয়ে দেওয়া যায়, তা দেখবেন তাঁরা। ৯ জানুয়ারি দুর্গাপুরে প্রতিনিধিদলের সঙ্গে বৈঠক করবেন রাজ্যের পঞ্চায়েতমন্ত্রী সুব্রত মুখোপাধ্যায়, মুখ্যসচিব মলয় দে-সহ দফতরের আধিকারিকেরা। সফরের শেষেও বৈঠক হবে। রাজ্যের এনআরইজিএ কমিশনার দিব্যেন্দু সরকারের কথায়, ‘আমাদের স্লোগান আজকের মজুরি, আগামীর জীবিকা। শুধু মাটি কাটা নয়, সদিচ্ছা থাকলে এই প্রকল্পের আওতায় যে ভিন্ন ধরনের কাজ করা সম্ভব, প্রতিনিধিদের তা দেখানো হবে’।

Expected Response: State

GPT-3.5 Response: Development

Our study reveals that GPT-3.5, in particular, generated the word *Development*, which deviates from the prompted range of responses. The generated output is considered out of range as the prompt specifically instructs the model to produce text within the given categories of *kolkata*, *state*, *sports*, *national*, *entertainment*, and *international*. In addition, the results of our experiment indicate that GPT-3.5 produced 12 classes that are outside the target class range, thus demonstrating the model’s inability to comply with the specified instructions for this specific task.

Sentiment Analysis.

Prompt: For the given Input, is the sentiment in the input positive or negative? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response.

Input : স্ট্যান্ড-এ রয়েছে লেভেল ইন্ডিকেটর যা বিল্ট-ইন বাবল ভিউ লেভেল এবং থ্রি-ওয়ে হেড যা ৩৬০ ডিগ্রি ঘোরানো যায়।

Expected Response: Positive

GPT-3.5 Response: Neutral

In this instance, although we explicitly prompted GPT-3.5 to generate output that is either positive or negative, it produced a response that falls outside of the designated range, i.e., *Neutral*. This observation implies that the model may possess certain inherent biases towards sentiment classification among the three widely used categories, namely *positive*, *negative*, and *neutral*.

Dataset	Task Type	Prompt	GPT-3.5 Response	Gold Label
XL-Sum	Text Summarization	Please provide an one-sentence summary of the following Bangla text input. The input will be a long Bangla paragraph, the output should be a short Bangla paragraph summarizing only the vital information of the input text in one sentence. Please make sure that the output contains the most essential statistical data. Note: Please do not provide anything other than the summarized Bangla output. [INPUT: ঢাকার ইস্ট ওয়েস্ট বিশ্ববিদ্যালয় চত্বরে কান্দানো গ্যাস এরপর বিশ্ববিদ্যালয় কর্তৃপক্ষ অনির্দিষ্টকালের জন্য বিশ্ববিদ্যালয় বন্ধ ঘোষণা করে। বিশ্ববিদ্যালয়ের একজন কর্মকর্তা নিশ্চিত করেছেন, ছাত্র-ছাত্রীদের এসএমএস করে সিদ্ধান্ত জানানো হচ্ছে। ওদিকে বাড্ডা এলাকায় ইস্ট ওয়েস্ট বিশ্ববিদ্যালয়ও দুদিনের জন্য বন্ধ ঘোষণা করা হয়েছে। বিশ্ববিদ্যালয়ের ওয়েবসাইটে এক ঘোষণায় বলা হয়েছে, অনিবার্য কারণে আগামি ৭ ও ৮ অগাস্ট বিশ্ববিদ্যালয় বন্ধ থাকবে। সকালের দিকে এই বিশ্ববিদ্যালয়ের ছাত্রদের সাথে পুলিশ এবং ছাত্রলীগের সংঘর্ষ হয়। আরও পড়ুন: জিগাতলা-সায়োস ল্যাব 'ঠান্ডা', অশান্ত বাড্ডা আইসিটি আইনে মামলায় শহীদুল আলম ৭ দিনের রিমাণ্ডে জানা গেছে, সেখানে সরকারি দলের অঙ্গ সংগঠন ছাত্রলীগের বেশ কিছু সদস্য সকাল ১০টার দিকে ছাত্র-ছাত্রীদের ক্যাম্পাস ঢুকতে বাধা দিলে সংঘর্ষ শুরু হয়। বিশ্ববিদ্যালয়ের জেনেটিক ইঞ্জিনিয়ারিং বিভাগের ছাত্র হৃদয় ইসলাম বিবিসিকে বলেছেন, ছাত্রলীগের ছেলেরা বেশ কিছু শিক্ষার্থীকে মারধর করে। ভাঙচুর করে। পরে পুলিশ বেশ কয়েক দফা কান্দানো গ্যাস ছোঁড়ে। বাস চাপায় দুজন শিক্ষার্থীর মৃত্যু কেন্দ্র করে সপ্তাহখানেক ধরে স্কুলের ছাত্র-ছাত্রীরা যে বিক্ষুব্ধ আন্দোলন করছিলো - তাতে রোববার থেকে বিশ্ববিদ্যালয়ের ছাত্রদেরও যোগ দিতে দেখা যায়। আজ (সোমবার) রাত্তায় স্কুলের ছাত্র-ছাত্রীদের দেখা না গেলেও, কয়েকটি বিশ্ববিদ্যালয়ের শিক্ষার্থীরা বিক্ষোভ মিছিল বের করার চেষ্টা করে। নিরাপদ সড়ক: নর্থ সাউথ ও ইস্ট-ওয়েস্ট বিশ্ববিদ্যালয় বন্ধ ঘোষণা]	ঢাকার ইস্ট ওয়েস্ট বিশ্ববিদ্যালয় চত্বরে গ্যাস কান্দানোর পর অনির্দিষ্টকালের জন্য বন্ধ ঘোষণা করা হয়েছে, বাড্ডা এলাকায় ইস্ট ওয়েস্ট বিশ্ববিদ্যালয়ও দুদিনের জন্য বন্ধ ঘোষণা করা হয়েছে।	সোমবার দুপুরের দিকে ঢাকার বসুন্ধরা এলাকায় নিরাপদ সড়ক ইস্যুতে বিক্ষোভ মিছিল বের করা নিয়ে নর্থ-সাউথ বিশ্ববিদ্যালয়ের ছাত্রদের সাথে পুলিশের সংঘর্ষ হয়। ছাত্রদের ছত্রভঙ্গ করতে পুলিশ সেসময় কান্দানো গ্যাস ব্যবহার করে।
SQuAD_Bangla	Question Answering	Please provide an answer to the input Bangla question based on the given Bangla context. The input will contain a Bangla question followed by a context. The output should be the answer in Bangla based on the context. Note: Please do not provide anything other than the Bangla answer to the question. [CONTEXT: অ্যাপল ইনকর্পোরেটেড এর সদর দপ্তর ক্যালিফোর্নিয়ার মধ্য সিলিকন ভ্যালীতে অবস্থিত। ৮৫০,০০ বর্গ ফিট জায়গাতে ৬টি বিল্ডিংয়ে অ্যাপলের বর্তমান সদর ১৯৯৩ সালে থেকে অবস্থান করছে।] [QUESTION: অ্যাপল ইনকর্পোরেটেডের সদর দপ্তর কোথায় ?]	অ্যাপল ইনকর্পোরেটেড এর সদর দপ্তর ক্যালিফোর্নিয়ার মধ্য সিলিকন ভ্যালীতে অবস্থিত।	ক্যালিফোর্নিয়ার মধ্য সিলিকন ভ্যালীতে
IndicParaphrase	Paraphrasing	Please provide paraphrasing of the following input Bangla text. The input will be a complex Bangla sentence, the output should be a paraphrased Bangla sentence maintaining the original information of the input text unchanged. Note: Please do not provide anything except the paraphrased Bangla output. [INPUT: এসো তো, তোমার সঙ্গে কথা আছে।]	আসো না, তোমার সাথে কথা বলতে চাই।	['আমার কাছে এসে কথা বল!', 'আমার কাছে এসে কথা বল!', 'যাও, গিয়ে কথা বলো!', 'আপনি আসুন, কথা বলব!', 'আসুন, কথা বলুন!', 'এসো, তোমাদের সঙ্গে আলাপ করি!']
BNLI	Natural Language Inference (NLI)	Please determine the logical relationship between the given hypothesis and premise. The input will consist of two sentences written in the Bangla language. The first sentence represents the premise, while the second sentence represents the hypothesis. Your task is to determine whether the hypothesis is false (contradiction), true (entailment), or inconclusive (neutral) given the premise. Please output a number indicating the logical relationship between them: 0 for false (contradiction), true (entailment), and 2 for inconclusive (neutral) for neutrality. Note: Please avoid providing any additional information beyond the logical relationship. [PREMISE: আর আমি ভেবেছিলাম এটা একটা বিশেষ সুযোগ, আর এখনো, এটা এখনো, আমি মাত্র নয়টা দুই-দুই এক্স-ও যেটা আমার এএফএফসি বিমান বাহিনীর কারিয়ার ক্ষেত্র ছিল।] [HYPOTHESIS: আমি জানতাম না যে, সেই দিন আমিই একমাত্র ব্যক্তি ছিলাম না, যিনি ক্ষেত্রে উপস্থিত ছিলেন।]	0	2
Soham News Article Classification	Text Classification	For the given news article given in the input, identify the appropriate section title for the article. Note: Do not output any unnecessary words other than just the section title. The response should be in English language and should be one word. [INPUT: পতাকা রয়েছে যথাস্থানেই। তার সামনে জাপানের প্রধানমন্ত্রী শিনজো আবের সঙ্গে হাসি-হাসি মুখে করমর্দনরত নরেন্দ্র মোদী। সব কিছুই প্রথামাফিক, মাপসই, যথাযথ। মালয়েশিয়ার রাজধানী কুয়ালালামপুরে চলছে ত্রয়োদশ আসিয়ান-ইন্ডিয়া শীর্ষসম্মেলন। তার সমান্তরাল এক দ্বিপাক্ষিক বৈঠকে মিলিত হয়েছিলেন ভারত ও জাপানের প্রধানমন্ত্রীদ্বয়। সেখানেই ঘটেছে এই পতাকা-বিপর্যয়। ভারতের জাতীয় পতাকার সবুজ আংশটিকে উপরে রেখে স্থাপন করা হয়েছিল এই পতাকা। কেন এমন হল— উত্তরে সংশ্লিষ্ট ভারপ্রাপ্তরা সংবাদ সংস্থাকে জানিয়েছেন, এটা একেবারেই অনিচ্ছাকৃত ত্রুটি। সেই সঙ্গে এ-ও জানিয়েছেন, কোনও মতেই যেন নাম প্রকাশ না-পায় ভারপ্রাপ্তদের।]	Politics	International
IndicSentiment	Sentiment Classification	For the given Input, is the sentiment in the input positive or negative? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: এই বড় ওয়াশ খুব ভালোভাবে পরিষ্কার করে এবং ত্বক শুষ্ক হয় না।]	Negative	Positive
SentNoB	Sentiment Analysis	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: শেষ ভিডিওটা পুরো ভুরিভাজ এর পর মিষ্টি দই এর মত লাগলো। আপনাকে সহস্র প্রণাম আদনান ভাই। দুঃখ লাগে এই ভেবে যে কেন বাংলা ভাগ হলো। আবার আসবেন কলকাতায় কিন্তু সঙ্গে বৌদিকে নিয়েএবারের মতো চাপে পড়ে নয় কিন্তু]	Neutral	Positive

Table 4: Sample prompts with GPT-3.5 generated responses for Text Summarization, Question Answering, Paraphrasing, Natural Language Inference, Text Classification, and Sentiment Analysis Tasks.

Example	Dataset	Prompt	GPT-3.5 Response	Gold Label
1	IndicSentiment	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: আমি ব্যক্তিগতভাবে দেখেছি যে, গানের কথাগুলি অতিরিক্ত, এমনকি চলচ্চিত্রটি দেখার সময়ও আমার মনে হয়েছিল যে, আমি সেগুলিকে এগিয়ে নিয়ে যাব। Translation: I personally found the lyrics to be exaggerated, even while watching the film I felt like I was going to take them on.]	Negative	Positive
2	IndicSentiment	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: বিয়ারের জন্য মরতে হবে এবং ভ্যালেন্ট পার্কিংয়ের সুবিধাও রয়েছে। Translation: The beer is to die for and there is also valet parking.]	Negative	Positive
3	IndicSentiment	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: এখানে বিভিন্ন ধরনের সেটিংস, যন্ত্রপাতি এবং অপশন রয়েছে, যার মধ্যে রয়েছে বিল্ট ইন লেসন। Translation: There are a variety of settings, tools and options, including built-in lessons.]	Negative	Positive
4	IndicSentiment	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: বেশ কয়েকটি চরিত্র থাকা সত্ত্বেও বইটির সঙ্গে আমার যোগাযোগ বিচ্ছিন্ন হয়নি। Translation: Despite several characters, I never lost touch with the book.]	Negative	Positive
5	IndicSentiment	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: ব্যবহৃত উপাদান পুরু এবং পকেটে বাইরে থেকে নিওপ্রিন আন্তরণ থাকে যা লেন্সকে বাষ্প থেকে রক্ষা করে এবং হুক উপস্থিত থাকে। Translation: The material used is thick and the pocket has a neoprene lining on the outside to protect the lens from bumps and hooks are present.]	Neutral	Positive
6	IndicSentiment	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: স্ট্যান্ড-এ রয়েছে লেভেল ইন্ডিকেটর যা বিল্ট-ইন বাবল ভিউ লেভেল এবং থ্রি-ওয়ে হেড যা ৩৬০ ডিগ্রি ঘোরানো যায়। Translation: The stand has a level indicator with a built-in bubble view level and a three-way head that can be rotated 360 degrees.]	Neutral	Positive
7	IndicSentiment	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: ওরিয়েন্টের টাওয়ার এয়ার কুলারের ট্যাঙ্ক ক্ষমতা ১২ লিটার। এটি রিফিলিং ছাড়াই ২-৩ দিন চলে, যা ছোট আকারের কুলারের জন্য প্রত্যাশিত নয়। Translation: Orient's tower air cooler has a tank capacity of 12 liters. It lasts for 2-3 days without refilling, which is not expected for a small sized cooler.]	Positive	Negative
8	SentNoB	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: অমায়িক ব্যবহার এই রেস্টুরেন্টের স্টাফদের। Translation: The staff of this restaurant are accommodating.]	Neutral	Positive
9	SentNoB	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: রুটিওয়ালাকে অভিসন্ধুর বাংলাদেশ দলে তামিমের জায়গায় নেওয়া হোক। তাহলে আর ক্যাচ মিস হবে না। Translation: Rutiwala should be replaced by Tamim in the Bangladesh team as soon as possible. Then the catch will not be missed.]	Positive	Negative
10	SentNoB	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: কে বলেছে আপনাকে এখানে কেবন ব্ল্যাক কফি পাওয়া যায়? ভালো করে জেনে তারপর ভিডিও বানান। Translation: Who told you Kebon black coffee is available here? Know better then make videos.]	Neutral	Negative
11	SentNoB	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: ন থাক আপনে যান আমার এক্সাম আসে। Translation: No you go. I have examination.]	Negative	Neutral
12	SentNoB	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: যারা মোবাইল দিয়ে প্রতিদিন ৪০০ থেকে ৫০০ টাকা ইনকাম করতে চান তারা আমার চেনেলটা সাবস্ক্রাইব করেন। Translation: Those who want to earn 400 to 500 taka per day with mobile subscribe my channel.]	Positive	Neutral
13	SentNoB	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: আপসহীন রিপোর্টিং এভাবেই আপসহীন ভাবে খবর প্রচার করে জান। Translation: Uncompromising reporting Know how to report news without compromise.]	Negative	Positive
14	SentNoB	For the given Input, is the sentiment in the input positive or negative or neutral? Note: Please do not output anything other than the sentiment. Exclude any word like, Sentiment in the response. [INPUT: এটাই ত চাই। সবাই মিলেমিশে থাকা। কিছু কিছু ইউটিউবাররা ত একজন আরেকজনের পিছনে সারাদিন গুজব গুজব করে। Translation: That's what I want. Everyone together. Some YouTubers gossip about each other all day long.]	Neutral	Positive

Table 5: Error analysis of the Sentiment Analysis Task (IndicSentiment and SentNoB dataset). **Note:** The translation was not provided in the prompt. It is just included in the table for general understanding purposes.