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 Nayeem[3], M Saiful Bari[4], Enamul Hoque[2]
[1]Islamic University of Technology, [2]York University, [3]University of Alberta,
[4]Nanyang Technological University, [5]Dialpad Inc.

* Equal contribution

Introduction

- We present BenLLM-Eval, an evaluation of LLMs to benchmark their performance in a modest resourced language, i.e., Bengali
- We evaluate three popular LLMs, i.e, GPT-3.5, LLaMA-2-13b-chat, and Claude-2 in zero-shot setting
- We carefully select seven important and diverse Bengali NLP tasks, i.e., text summarization, question-answering, paraphrasing, natural language inference, transliteration, text classification, and sentiment analysis
- Experimental results suggest in most of the tasks, their performance is moderate (with LLaMA-2-13b-chat performing significantly bad) in comparison to state-of-the-art results

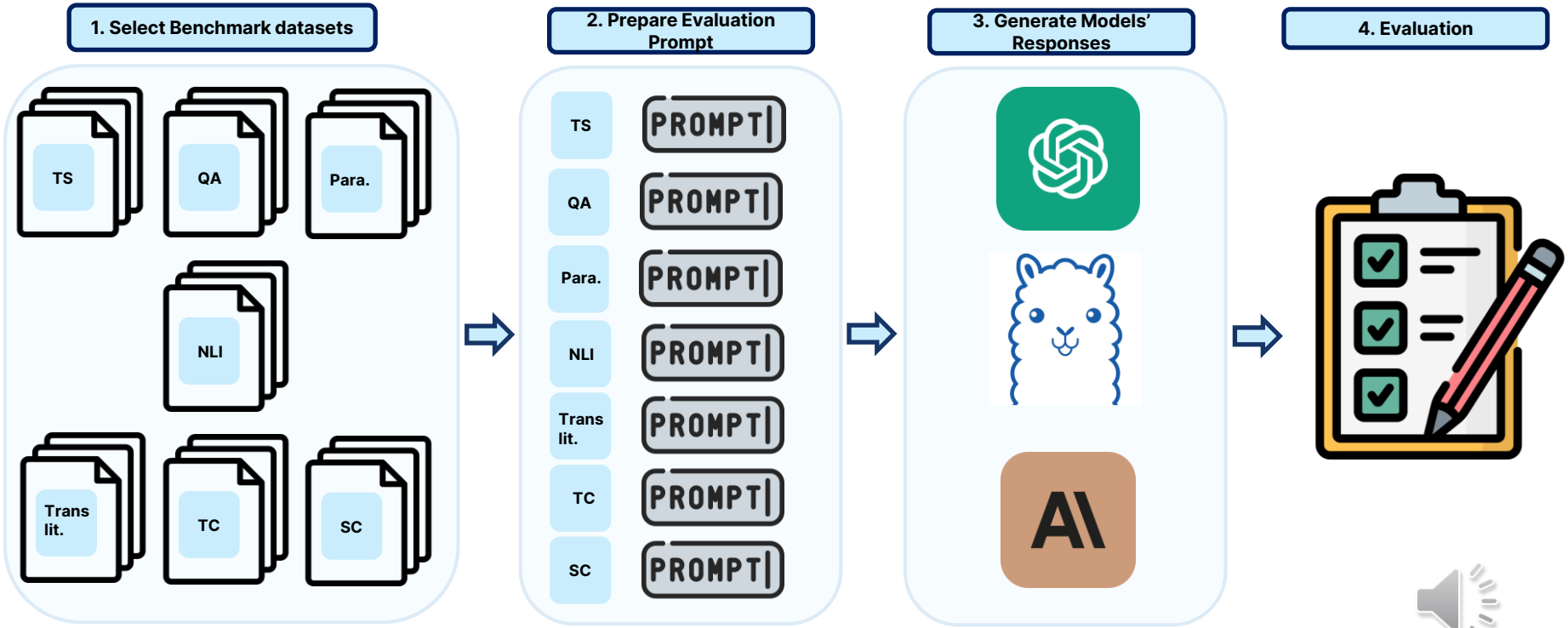
Motivation

- Despite the impressive capabilities of LLMs, they may still frequently generate untruthful facts that diverge from the original input
- Furthermore, ChatGPT like LLMs has demonstrated strong zero-shot performance in various NLP tasks in English and some other languages and domains,
- Yet these LLMs are to be investigated in the widely spoken, yet modest-resourced, Bengali language domain

Contributions

- To our knowledge, our work is the first to evaluate three popular LLMs, i.e, GPT-3.5, LLaMA-2-13b-chat and Claude-2 in zero-shot setting
- We evaluate the performance of the LLMs in seven benchmark tasks:
  - Text Summarization → 1 dataset
  - Question-Answering → 1 dataset
  - Paraphrasing → 1 dataset
  - Natural Language Inference → 1 dataset
  - Transliteration → 1 dataset
  - Text Classification → 1 dataset
  - Sentiment Analysis → 2 datasets
- We also perform task contamination analyses which helps to identify a model's prior exposure to test tasks on its training data
- We share the LLM-generated responses, prompts, and parsing scripts for all seven tasks

Methodology



Datasets and Prompts

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:]

Results

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**.

Results

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**.

Results

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.

Results

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.

Results Analysis

- While in most tasks ChatGPT-3.5 and Claude-2 performed moderately, they performed well in the sentiment analysis task compared to the SoTA results
- However, in all of the tasks, the performance of the LLaMA-2-13b-chat model was significantly poor
- In the transliteration task, ChatGPT-3.5 was the best performer

Task Contamination Analysis

- Task contamination analysis is essential to ensure a fair model evaluation since it helps identify a model's prior exposure to test tasks on its training data
- 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

Task Contamination Analysis Results

- Our findings reveal that only GPT-3.5 could generate examples related to the tasks (Sentiment Analysis, Text classification 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

Task Contamination Analysis Results

- 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
- 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

Conclusion and Future works

- We introduce BenLLM-Eval, which provides a comprehensive zero-shot evaluation of LLMs on seven benchmark NLP tasks
- The results revealed that 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
- 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
- In the future, we intend to expand our experiments by including additional low to modest-resource languages, tasks, datasets, and settings

THANK YOU