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

Mohsinul Kabir[1], Mohammad Saidul Islam[2], Md Tahmid Rahman Laskar[2], 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

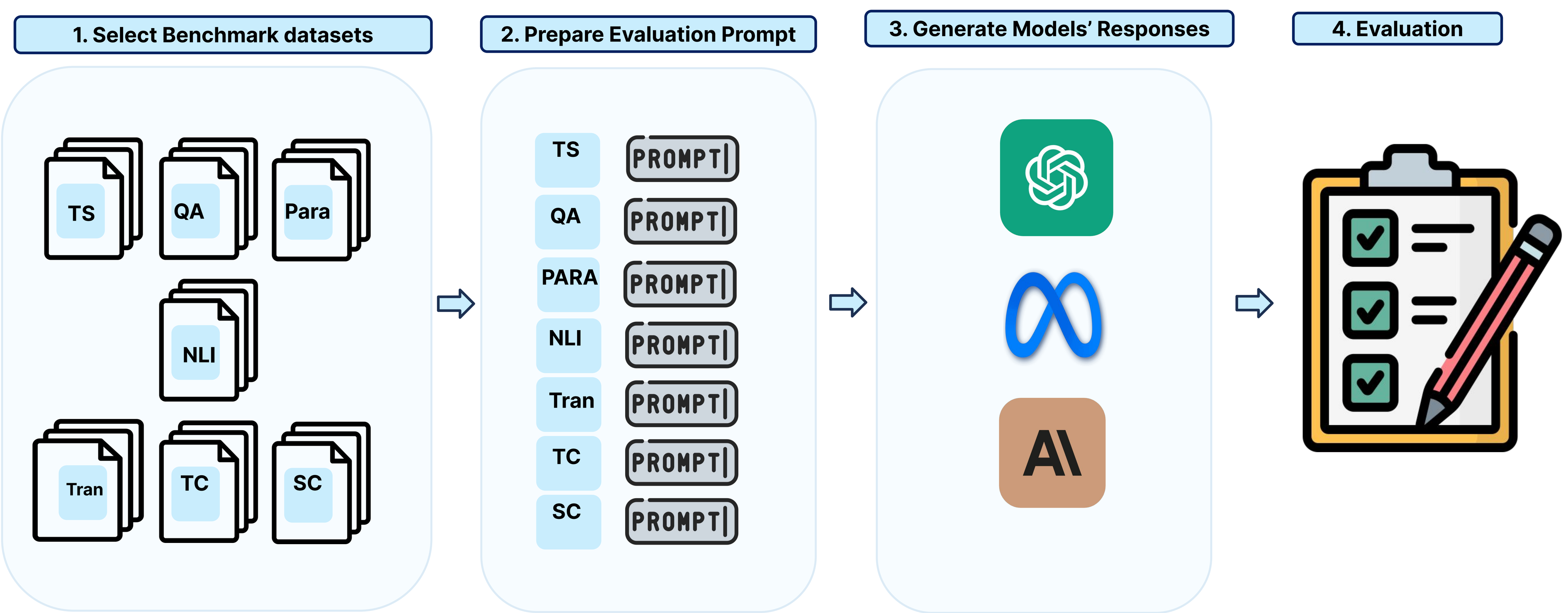
Introduction

We present BenLLM-Eval, an evaluation of LLMs to benchmark performance in a modest-resourced Bengali language. We evaluate 3 LLMs, namely, GPT-3.5, LLaMA-2-13b-chat, and Claude-2 in zero-shot setting. We select seven diverse Bengali NLP tasks, namely, text summarization, question-answering, paraphrasing, natural language inference, transliteration, text classification, and sentiment analysis

Motivation and Contributions

- No prior work evaluated GPT-3.5, Claude-2 and Llama-2 on Bengali NLP tasks.

Methodology



Task and Datasets

- We evaluate the performance on 7 benchmark Bengali NLP tasks (see Table 1 for more details):
  - **Text Summarization:** 1 dataset (XL-Sum)
  - **Question-Answering:** 1 dataset (SQuAD-Bangla)
  - **Paraphrasing:** 1 dataset (IndicParaphrase)
  - **Natural Language Inference:** 1 dataset (BNLI)
  - **Transliteration:** 1 dataset (Dakshina)
  - **Text Classification:** 1 dataset (Soham News Article)
  - **Sentiment Analysis:** 2 datasets (IndicSentiment & SenNoB)

Results and Discussion

- We demonstrate our results in Table 1 and Table 2.
- While in most tasks GPT-3.5 and Claude-2 performed moderately, they **performed on par** with the SoTA models in the **sentiment analysis** task.
- However, in all of the tasks, the performance of the **LLaMA-2-13b-chat** model was **significantly poor**.
- In the **transliteration** task, **GPT-3.5** was the **best** performer.

Task and Data Contamination Analysis

- We apply two contamination detection technique:
  - **Task Example Extraction.**
  - **Membership Inference.**
- Our findings reveal that only GPT-3.5 could generate examples related to tasks like **Sentiment Analysis, Text classification, Transliteration 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
- 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. However, we did not observe any exact match in **summarization**.
- In summary, **contamination could be an issue** with the **GPT-3.5 model in Sentiment Analysis**, **Text Classification,** 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.**

Conclusions and Future Work

- We present a comprehensive zero-shot evaluation of LLMs on 7 benchmark NLP tasks.
- Our results reveal that in some tasks, GPT-3.5 or Claude-2 perform on par (e.g., summarization) or even outperform (e.g., sentiment analysis) current SOTA models.
- In the future, we will expand our experiments by including more low to modest-resource languages, tasks, datasets, and settings

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 1: Performance Comparison between zero-shot LLMs and fine-tuned SOTA models on Text Summarization (TS), Question-Answering (QA), Paraphrasing (PP), Natural Language Inference (NLI), Text Classification (TC), and Sentiment Analysis (SA).

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 2: Single-word and Full-sentence transliteration results.