Revision

AE Comment:

Weakness: The primary concern of the reviewers lies in the limited contribution of the paper. Reviewers note that the paper is heavy on background, and suggest the short paper format for the paper's technical contributions instead. The paper focuses on a limited set of languages which are similar linguistically, it doesn't provide broad insight on low-resource languages in general. Moreover, there are several benchmarks for low-resource languages from around the world (reviewers noted the XTREME series of benchmarks, but there are several more including the UD treebank which contains annotated data in 100+ languages) – these existing benchmarks could be used in the paper to comprehensively test LLM abilities on these languages. The paper does not need to rely on translated data, which is not ideal for evaluation since translation quality is not quantified.

Revision: We address the AE comments and submit it as a short paper describing the technical contributions. We also address the differences between our work and existing benchmarks that focus on low-resource languages. Moreover, we also provide some insights that some studies use automatic translation for evaluation, particularly for South Asian languages.

Reviewer 1:

Weakness-1: There already exist multilingual benchmarks on which LLMs have been evaluated (such as BUFFET: https://arxiv.org/pdf/2305.14857.pdf). In addition, there are other multilingual benchmarks that have been used by the community (such as XTREME, and XTREME-R). These benchmarks include the languages/tasks the authors have considered. These benchmarks are not even cited in this paper. And hence, it is not justified why a new benchmark is interesting.

Revision: We incorporated the reviewer's comment in this version by discussing the multilingual benchmarks and how these benchmarks are different from our work. While existing multilingual benchmarks investigate few-shot and fine-tuning using smaller LLMs giving minimal focus on South Asian languages, we opt to choose zero-shot learning on state-of-the-art LLMs concentrating on the most spoken South Asian languages only.

Weakness-2: The authors report that the translated datasets were of moderate quality and many hashtags were not even translated. This brings into question the quality of these datasets and their usefulness for the general community. I also wonder if this dilutes the result because the evaluated models could rely on English hashtags to make predictions.

Revision: We clarify the translation quality and provide examples of similar studies that use automatic translation, particularly for South Asian languages (Appendix C.1.1). We also clarify the reason for not removing hashtags and LLMs do not solely rely on hashtags but on the entire sequence (Section 4).

Weakness-3: Minor, but the paper takes too long to get to the main content. The related work and the background section do not add to understanding the main content of the paper. I would suggest the authors to make this a short paper in the next iteration but trimming the findings to 4 pages.

Revision: We adopted the reviewer's suggestion and trimmed the findings to 4 pages.

Reviewer 2:

Weakness-1: The issue of LLMs performing poorly on low-resource languages compared to English is already well-known within the NLP community.

Revision: We clarify the addressed point in this version. We also want to add that Gemini highlights superior performance on low-resource languages in sentiment tasks. Moreover, despite not supporting Urdu, Gemini still exhibits promising results, sometimes outperforming Bangla. Our focus on the most spoken South Asian languages, including Hindi, Bangla, and Urdu, underscores their significance despite limited computational resources.

Weakness-2: The paper leaves some ambiguity regarding the novel insights or practical implications derived from this study, raising questions about the extent to which it advances our understanding or solutions in addressing the performance gap for low-resource languages.

Revision: We addressed the comment in this version providing more information on novel insights and practical implications derived from our study.

Comment-1: Given the long form, I would have liked to see some qualitative analysis with examples. Perhaps this would shed light on how LLMs can be improved for Bengali, Hindi, and Urdu in NLI, Sentiment Analysis, and Hate Speech Detection.

Revision: We changed our submission to a short paper based on the AE and reviewers' comments.

Reviewer 3:

Weakness-1: LLM performances on different languages are not completely balanced. This conclusion seems obvious and has been verified by a large amount of work[1-4]. Among them, the XNLI performance has been verified by [3,4].

Revision: We add more details on how our work differs from existing benchmarks in this submission. Moreover, our study uniquely focuses on comparing resource-rich (English) and low-resource (Bangla, Hindi, and Urdu) languages using state-of-the-art LLMs.

Weakness-2: Missing Reference [3,4]

Revision: We addressed the comment in this version adding the references.

Weakness-3: I think that direct translation is likely to loss the sentiment analysis information, resulting in a smaller gap between different languages in Gemini. And the translation quality is also said to be moderate in the article. Is the conclusion of this benchmark credible?

Revision: We add a discussion on direct translation is less likely to loss the sentiment information based on the smaller performance gap in Gemini, significant disparities persist in GPT-4 and Llama-2. Moreover, we also discussed the translation quality and provided some existing studies that use a similar approach.

Weakness-4: I think the translation of Hate Speech is even more unreasonable, because this kind of moral or social issue itself has a strong social background. For example, saying someone looks like a giraffe is offensive in the English context, but in Chinese or Japanese context, it has a neutral or joking nature.

Revision: We clarified the translation quality in this updated version. Moreover, Our methodology aligns with established practices from previous studies [1, 2] that translate reputable datasets [3-5].

[1] INDICXNLI: Evaluating Multilingual Inference for Indian Languages (https://aclanthology.org/2022.emnlp-main.755.pdf)
[2] Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback (https://aclanthology.org/2023.emnlp-demo.28.pdf)
[3] https://huggingface.co/datasets/jon-tow/okapi_hellaswag
[4] https://huggingface.co/datasets/jon-tow/okapi_arc_challenge
[5] https://huggingface.co/datasets/jon-tow/okapi_truthfulqa

Weakness-5: Except for XNLI, it does not provide more rigorous, language-independent general tasks such as mathematics to illustrate explicit gaps between different languages.

Revision: While our current study focused on NLI, Sentiment, and Hate Speech tasks for South Asian languages compared to English, we appreciate the suggestion to include more rigorous, language-independent tasks like mathematics in future research.

Comment-1: Missing Reference [1,2]. Please add more references for multilingual LLMs.

Revision: We added above mentioned references along with more multilingual benchmarks in this version.

Comment-2: Please explain in detail why these languages were chosen as low-resource languages, and why not expand to more low-resource languages?

Revision: We addressed this comment and added a discussion for choosing these languages as low-resource languages.

Comment-3: Please add a discussion of the low-resource language, whether it is a low-resource language for human society or a low-resource language for the model. Because for LLaMA, its vocabulary does not support non-Latin languages, like Chinese. Therefore, all non-Latin languages can become low-resource languages for For LLaMA.

Revision: We added a discussion in this version that our classification of Bangla, Hindi, and Urdu as low-resource languages pertains to their scarcity in available datasets rather than limitations in LLM capabilities.