Large Language Models As Annotators: A Preliminary Evaluation For Annotating Low-Resource Language Content

Savita Bhat[†‡] and **Vasudeva Varma**[‡]
[†]TCS Research, [‡]IIIT Hyderabad
savita.bhat@tcs.com, vv@iiit.ac.in

Abstract

The process of collecting human-generated annotations is time-consuming and resource-hungry. In the case of low-resource (LR) languages such as Indic languages, these efforts are more expensive due to the dearth of data and human experts. Considering their importance in solving downstream applications, there have been concentrated efforts exploring alternatives for human-generated annotations. To that extent, we seek to evaluate multilingual large language models (LLMs) for their potential to substitute or aid human-generated annotation efforts. We use LLMs to re-label publicly available datasets in LR languages for the tasks of natural language inference, sentiment analysis, and news classification. We compare these annotations with existing ground truth labels to analyze the efficacy of using LLMs for annotation tasks. We observe that the performance of these LLMs varies substantially across different tasks and languages. The results show that off-the-shelf use of multilingual LLMs is not appropriate and results in poor performance in two of the three tasks.

1 Introduction

Traditionally, compiling annotations using human experts has been the primary step in formulating a supervised solution[1] for various tasks such as sentiment analysis (Rosenthal et al., 2017), bot detection (Fagni et al., 2021), and inference (Bowman et al., 2015; Wang et al., 2018). The process of collecting human-generated annotations is often time-intensive and resource-hungry. Specifically, in the case of LR languages, these efforts are more expensive due to a lack of quality data and human experts. Therefore, alternatives to human-generated labels are being actively explored (Cruz and Cheng, 2020; Magueresse et al., 2020).

Recent LLMs[2], such as *ChatGPT*, demonstrate impressive performance in various NLP applications such as summarization, classification, and text generation (Liu et al., 2023). Furthermore, interesting use cases and applications using these generative models have been explored and reported (Zhao et al., 2023). The research community is curious to know how close LLMs are to human experts and annotators. Accordingly, (Guo et al., 2023) conduct extensive evaluations in a question-answering setup. In (Zhu et al., 2023), *ChatGPT* is evaluated in the context of reproducing human-generated label annotations in social computing tasks. Similar studies for misinformation in (Bang et al., 2023) and hate speech in (Huang et al., 2023) have considered *ChatGPT* for annotations. Additionally, several works (Kuzman et al., 2023; Gao et al., 2023; Wang et al., 2023) compare *ChatGPT*'s annotation and evaluation performance with human experts.

The point to note is that most of these efforts focus on high-resource (HR) languages like English. In reality, these HR languages are not recognized as the native languages for most of the world's population. For example, people in India prefer to interact in one of the Indic languages despite of being literate in English. These Indic languages are generally categorized as low-resource (LR) languages because of the unavailability of quality data sources (Lai et al., 2023). Considering India as the most populated country[3] in the world, it is essential to evaluate current multilingual LLMs in the context of LR languages like Indic languages. Secondly, besides *ChatGPT*, other multilingual LLMs like *mT0* and *BLOOMZ* must also be evaluated for such use cases.

To this extent, we primarily explore the possibility of using multilingual LLMs as a substitute for human annotators. Specifically, we focus on low-resource languages such as Indic languages and compare the LLM-generated annotations with the

---

[1] https://en.wikipedia.org/wiki/Supervised_learning

[2] LLMs and generative models are used interchangeably.

[3] https://tinyurl.com/2tz9d3u2; Last accessed: 09/06/2023

Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, pages 100–107
November 1, 2023 ©2023 Association for Computational Linguistics

ground truth human-generated labels. To the best of our knowledge, this is the first work to evaluate the efficacy of LLMs as annotators for LR Indic languages. We examine three LLMs- ***ChatGPT***, ***mT0***, and ***BLOOMZ***, for three tasks- document classification, sentiment analysis, and natural language inference. The main observations from our experiments are as follows:

1. All three LLMs perform well in identifying sentiments. Surprisingly, *ChatGPT* shows slightly worse capability for simple classification, parsing, and inference tasks. It does remarkably well in a more complex task of news category classification.
2. The performance of these LLMs, in correctly annotating the samples, is not uniform and varies across different tasks and different LR languages. This observation demands more informative, clear, and better prompts/instructions while using generative models as annotators.
3. Fine-tuned baseline models have superior performance in most of the languages and tasks, highlighting the need for focused task-specific training.
4. *ChatGPT* is the only LLM that often provides a justification with the answer, which helps in understanding annotation choices.

2 Methodology

We follow a comparative approach to study the differences between human-generated and LLM-generated annotations for Indic languages. Under this premise, we consider three broad categories of tasks and relevant datasets: 1) **WNLI** - Winograd inference task involving inference based on a given context, 2) **SA** - identifying sentiment for a given text, and 3) **NewsCLS** - categorizing given news text. We consider appropriate prompting strategies to simulate the manual annotation process. In the following subsections, we describe the multilingual LLMs used for annotations (Section 2.1), Datasets used for the three tasks (Section 2.2), and our approach for the annotation process (Section 2.3).

2.1 LLMs

We explore the following LLMs in the context of Indic languages for our annotation experiments. The choice of LLMs was guided by the following constraints: 1) LLM should be trained on multilingual data sources, including Indic languages, and 2) LLM training consists of multiple tasks converted to text-to-text format. This way, we make sure that the strategies, i.e., the instructions to the selected LLMs, do not have large variations and are similar in nature.

ChatGPT (GPT-3.5) is known to be created by finetuning the GPT-3.5 variant using reinforcement learning from human feedback (**RLHF**) (Christiano et al., 2017). We evaluate this model using *gpt-3.5-turbo* API between 5th September to 6th September 2023. Even though there is no definite information released by OpenAI on this model, it is assumed that'CommonCrawl' corpus, which contains some percentage of data in Indic languages, is a part of the training data for this model [4].

BLOOMZ (Muennighoff et al., 2022) is an open-source multilingual LLM. Multitask prompted fine-tuning (MTF) is applied to pretrained BLOOM LLM (Scao et al., 2022) to build the fine-tuned variant, *BLOOMZ*. *BLOOMZ* family consists of models with 300M to 176B parameters and supports 59 languages.

mT0 (Muennighoff et al., 2022) is the fine-tuned variant of pretrained multilingual mT5 language model. Like *BLOOMZ*, MTF is applied to mT5 to produce *mT0* with model variants ranging from 300M to 176B.

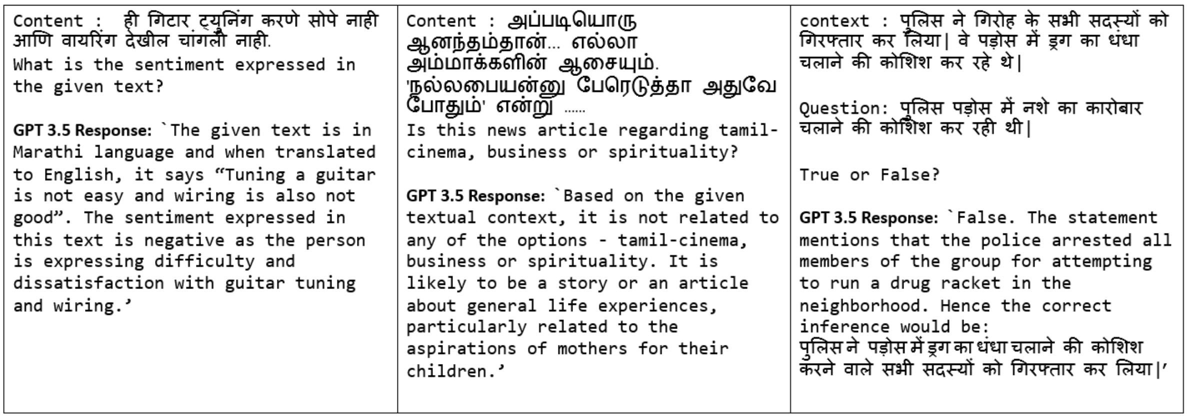
BLOOMZ and *mT0* families have been trained on datasets, xP3 and xP3MT, consisting of 13 training tasks in 46 languages. xP3 uses English prompts, whereas xP3MT uses prompts machine-translated to 20 languages. Indic languages constitute a small part of the training data for both of these model families.

2.2 Datasets

We consider 11 Indic languages as LR languages for our experiments. It should be noted that not all of these languages have quality datasets identified and compiled for certain tasks. We choose the datasets and tasks with maximum representation from Indic languages. Out of 11 Indic languages, Hindi is the only medium-resource (MR) language, whereas Punjabi, Oria, and Assamese are classified as extremely low-resource (XR) languages. The remaining Indic languages are categorized as LR

---

[4] https://en.wikipedia.org/wiki/GPT-3. *Last accessed on* 6th September 2023



Content : ही गिटार ट्युनिंग करणे सोपे नाही आणि वायरिंग देखील चांगली नाही.
What is the sentiment expressed in the given text?

GPT 3.5 Response: `The given text is in Marathi language and when translated to English, it says “Tuning a guitar is not easy and wiring is also not good”. The sentiment expressed in this text is negative as the person is expressing difficulty and dissatisfaction with guitar tuning and wiring.’

SA - Marathi

Content : அப்படியொரு ஆனந்தம்தான்... எல்லா அம்மாக்களின் ஆசையும். 'நல்லபையன்னு பேரெடுத்தா அதுவே போதும்' என்று ......
Is this news article regarding tamil-cinema, business or spirituality?

GPT 3.5 Response: `Based on the given textual context, it is not related to any of the options - tamil-cinema, business or spirituality. It is likely to be a story or an article about general life experiences, particularly related to the aspirations of mothers for their children.’

NewsCLS - Tamil

context : पुलिस ने गिरोह के सभी सदस्यों को गिरफ्तार कर लिया| वे पड़ोस में ड्रग का धंधा चलाने की कोशिश कर रहे थे|

Question: पुलिस पड़ोस में नशे का कारोबार चलाने की कोशिश कर रही थी|

True or False?

GPT 3.5 Response: `False. The statement mentions that the police arrested all members of the group for attempting to run a drug racket in the neighborhood. Hence the correct inference would be:
पुलिस ने पड़ोस में ड्रग का धंधा चलाने की कोशिश करने वाले सभी सदस्यों को गिरफ्तार कर लिया|’

WNLI - Hindi

Figure 1: Prompt templates for annotation. *ChatGPT* responses show the justifications provided for annotation choices. *mT0* and *BLOOMZ* do not provide any justification.

languages (Lai et al., 2023). This categorization is based on their representation in the CommonCrawl corpus. **IndicNLPSuite** (Kakwani et al., 2020) introduced NLP resources for Indic languages. We choose the following datasets from its **IndicGLUE** evaluation benchmark. The datasets are selected based on two criteria, 1) the datasets are annotated by humans, and 2) the dataset covers as many of Indic languages as possible.

Sentiment Analysis (SA) We use **IndicSentiment**[5] dataset from *Huggging Face* datasets. Each example contains a review text and corresponding sentiment. As per the dataset card, the annotations are expert-generated. The input records in the dataset are translated into various Indic languages (Doddapaneni et al., 2023). The task is to identify the sentiment of a given text.

News Category Classification (NewsCLS) The task is to categorize a news article into a given set of topics. This dataset is compiled by crawling regional news websites. We assume that the categories are manually assigned to the news articles based on the URLs while publishing on the website.

Winograd NLI (WNLI) We use the Indic version of WNLI dataset (Kakwani et al., 2020). The dataset is created and verified by experts by translating the original dataset into 3 Indic languages (mr, hi, gu). Each example consists of a pair of sentences where the second sentence is constructed from the first sentence by replacing an ambiguous pronoun with a possible referent within the sentence. The task is to predict if the original sentence entails the second sentence.

2.3 Annotation

We attempt to re-annotate the data samples for each task and dataset using *ChatGPT*, *BLOOMZ*, and *mT0*. We use PromptSource toolkit (Bach et al., 2022) to identify candidate prompts for our tasks. We experiment with relevant prompts and choose the ones appropriate for chosen LLMs and tasks. Although the context is given in Indic languages, the prompts are in English. Example prompts are presented in Figure 1.

SA For the SA task, we ask the LLMs to identify the for a given context as follows:
`Content: {text content}`
`What is the sentiment expressed in the given text?`
where `{text content}` is the review text in a LR language.

NewsCLS This task consists of categorizing given news content in one of the categories. It is observed that the news records in every language have a certain closed set of categories. We use these sets to modify the prompt template as below:
`Content: {news content} Is this news article regarding {categories}?`
where `{news content}` is the news text and `{categories}` is the set of candidate categories.

---

[5] https://huggingface.co/datasets/ai4bharat/IndicSentiment

Task	Language	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te
SA	mT0	0.910	0.915	0.911	0.931	0.911	0.898	0.929	0.763	0.856	**0.947**	0.890
	BLOOMZ	0.927	**0.955**	**0.944**	**0.971**	0.899	**0.939**	**0.942**	**0.938**	0.927	0.940	0.891
	ChatGPT	0.856	0.8761	0.845	0.909	0.839	0.843	0.836	0.772	0.846	0.822	0.768
	mBERT	0.57	0.68	0.66	0.73	0.68	0.68	0.69	0.49	0.75	0.71	0.66
	indicBERTplus	**0.931**	0.93	0.933	0.933	**0.928**	0.932	0.938	0.931	**0.933**	0.936	**0.937**

Table 1: Sentiment Analysis: Language-wise weighted F1-score for ***mT0***, ***BLOOMZ***, and ***ChatGPT***. The bold number indicates the highest value per language, whereas the red colour denotes the highest performance amongst multilingual LLMs for every language.

WNLI Since the task is to identify entailment given a context and secondary sentence, we consider the prompt where the entailment is explored through a true/false question. The prompt used is as follows:

```
Context: {sentence1}
Question: {sentence2}
True or False?
```

where `{sentence1}` and `{sentence2}` are the context and secondary sentence respectively.

3 Experimental Setup

As mentioned earlier, we primarily use three tasks and corresponding datasets to evaluate if LLMs can replace or to some extent, aid the manual annotation efforts. We formulate the annotation task as a zero-shot inference task. We compare LLM annotations with the ground truth labels. We consider '*test*' split from all the datasets to ensure no data leakage. The '*gpt-3.5-turbo*' API for *ChatGPT* is paid and under a constrained usage policy. Hence, we use a subset of samples for the *ChatGPT* experiments. For *mT0* and *BLOOMZ*, we use the entire split whenever possible. The dataset distributions are as follows: We use the entire '*test*' split distributed across various Indic languages for WNLI and NewsCLS tasks, totaling to 284 and 5986 data samples, respectively. For sentiment analysis, we randomly select a total of 2862 samples spread across 11 languages with approximately 250 samples each, considering the budget for the paid experiments with the '*gpt-3.5-turbo*' API. We use the following abbreviations for languages: as (Assamese), bn (Bengali), gu (Gujarati), hi (Hindi), kn (Kannada), ml (Malayalam), mr (Marathi), or (Odia), pa (Punjabi), ta (Tamil), and te (Telugu).

For *ChatGPT*, we use the official OpenAI API (*gpt-3.5-turbo*) with default settings to annotate the samples. Similarly, we use Hugging Face models and tokenizers for *mT0* and *BLOOMZ* LLMs for annotations. Due to infrastructure constraints, we use the '*mT0-large*' model for *mT0* and the '*BLOOMZ-1b1*' model for *BLOOMZ* experiments. No training is involved since we consider zero-shot inferencing with the off-the-shelf model, i.e., a zero-shot setting. For comparison, we consider state-of-the-art baselines finetuned for these specific tasks. For Sentiment Analysis, we use results reported in (Doddapaneni et al., 2023), while results from (Kakwani et al., 2020) are considered as baseline for WNLI and NewsCLS tasks.

We use weighted-precision, weighted-recall, and weighted-F1 metrics from *sklearn* library for evaluation. We also report macro-average calculated across all languages to indicate the correctness of labels for a specific task.

4 Results & Analysis

This section presents the overview of the annotation experiments for three tasks and three LLMs. Representative detailed language-wise performance results (F1 measure) for each task are listed in Table 1,2, and 3. Table 4 describes correctly labeled instances across different tasks and LLMs.

SA - Superior performance in zero-shot inference All three LLMs perform well in identifying sentiment for a given textual content. It is interesting to see that *ChatGPT* is ranked last amongst LLMs in most cases. In 9 out of 11 languages, *BLOOMZ* shows superior or at-par performance as compared to baseline models. It is encouraging to see good zero-shot inference with just a single instruction. We expect even better results with more informed and aligned prompting strategies.

SA - Additional information and justification It should be noted that *mT0* and *BLOOMZ* consider two sentiments (*Positive* and *Negative*) as candidates for the assignment. In contrast, *ChatGPT* considers three sentiments by default (*Neutral* as additional sentiment). After manual validation, we observe that the records are indeed of neutral sentiment. Secondly, *ChatGPT* also provides reasonable justification for the suggested annotations. These justifications are useful in providing clear instructions for training the crowd-workers for annotations. We believe that this additional information and justifications will help in aligning expert-generated and machine-generated annotations.

NewsCLS - Complex tasks need focused training and instructions To introduce more complexity in the classification task, we consider News category classification task. This is a multi-class problem with a very fuzzy class separation. Out of three LLMs, *ChatGPT* performs better in 6 out of 7 languages. The othe two LLMs demonstrate varied performance ranging from low to high accuracy. Additionally, all these LLMs lags behind the baselines and fail to reproduce the human-generated annotations. As can be seen the task-specific finetuning boosts the model performance. We believe that the prompts/instructions given to the LLMs were simple and unable to fully specify the complexity and requirements of the task. Accordingly, we conclude that complex tasks need more focused and aligned instructions to help the LLM in annotations.

NewsCLS - Appropriate corrections for noisy data samples We note that the annotations for this dataset are noisy, and a few records can be assigned to multiple categories instead of just one category. We believe that this may have affected the evaluation using automatic metrics. It is also observed that only *ChatGPT* looks beyond the candidate categories and suggests appropriate alternate categories that are valuable in annotation efforts. On manual validation, we observe that these suggestions are indeed relevant and useful.

WNLI - Reasoning and inference tasks are harder All models, including the three LLMs and baselines, show average performance in recreating the annotations for the inferencing task. It is interesting to note that the zero-shot inferencing with multilingual LLMs comes close to the performance of finetuned baseline models. In general, the reasoning and inferencing tasks require natural language understanding and hence are more complex to train for. With LR languages, the problem becomes harder, considering the unavailability of training and annotation resources. We believe that clear prompts and supplementary explanations will help in improving the performance.

WNLI - Justification may help in language understanding It is observed that only *ChatGPT* provides relevant justification for the inference in most cases. These justifications often explain the decision and the logical reasoning behind that decision. These justifications are useful in understanding the annotation choices and, hence, can serve as a guiding tool for better annotation alignment.

Annotation Correctness Percentages of correctly labeled samples for the three tasks and three LLMs are listed in Table 4. This is the macro average across all relevant languages for a particular task. It is interesting to see that *ChatGPT* performs far worse than *mT0* and *BLOOMZ* in the relatively simpler task of sentiment analysis. In NewsCLS, all three LLMs have poor showing, whereas in WNLI, only *ChatGPT* seems to have more than a chance performance. In the case of *mT0* and *BLOOMZ*, it is difficult to conclude that the performance is not random. The performance in individual languages documented in Table 3 does not seem to be a by-chance result. However, further investigation with more samples and varied prompts is required to understand this result.

LLMs for LR languages As mentioned earlier, LR languages occupy a small portion of the CommonCrawl campus. Consequently, the LLMs trained on this corpus also have a similar small representation in their embeddings, often demonstrating a limited linguistic understanding of these languages. It is reiterated by the F1 score and the correctly labeled portion in NewsCLS and WNLI tasks. These tasks require a certain degree of language understanding and reasoning capability, which none of the three LLMs demonstrate in any Indic language except Hindi.

Language families such as *Dravidian* (Kannada, Tamil, Telugu, and Malayalam) and *Indo-Aryan* (Hindi and Marathi) share a lot of commonalities among themselves. Despite that, the significant difference in the scores supports the dependence on language exposure during training. As can be seen from the results, the LLMs have different levels of understanding of these languages, and there seems to be no clear winner.

Annotations & Justifications We observe annotations provided by three LLMs, *mT0*, *BLOOMZ*, and *ChatGPT*. Only *ChatGPT* offers a justification while providing an answer/annotation. These justifications often explain the reasoning behind var-

Task	Language	bn	gu	hi	ml	mr	ta	te
NewsCLS	mT0	0.20	0.69	0.076	0.739	0.257	0.27	0.292
	BLOOMZ	0.26	0.69	0.18	0.6250	0.32	0.488	0.426
	ChatGPT	0.472	0.757	0.53	0.68	0.522	0.49	0.10
	mBERT	**0.80**	0.89	0.60	0.82	0.87	0.92	-
	IndicBERT	0.78	**0.92**	**0.74**	**0.94**	**0.94**	**0.96**	-

Table 2: NewsCLS Task: Language-wise weighted F1-score for ***mT0***, ***BLOOMZ***, and ***ChatGPT***. The bold number indicates the highest value per language and the red colour denotes the highest performance value amongst the multilingual LLMs for every language.

Task	Language	gu	hi	mr
WNLI	mT0	0.400	0.415	0.344
	BLOOMZ	0.3751	0.508	0.539
	ChatGPT	0.406	0.406	0.406
	mBERT	**0.56**	**0.56**	**0.56**
	IndicBERT	**0.56**	**0.56**	**0.56**

Table 3: WNLI Task: Language-wise weighted F1-score for ***mT0***, ***BLOOMZ***, and ***ChatGPT***. The bold number indicates the highest value for every language and the red color denotes the highest performance value amongst the multilingual LLMs for every language.

Task	mT0	BLOOMZ	ChatGPT
SA	89.8%	**93.4%**	83.8%
NewsCLS	32.1%	38.9%	**51.4%**
WNLI	38.8%	**47.7%**	40.6%

Table 4: Correctly labelled records by ***mT0***, ***BLOOMZ***, and ***ChatGPT***. The number in bold indicates the average highest performance for the corresponding task. We consider the macro-average calculated across all the relevant languages for a task.

ious annotation choices. We concur with (Huang et al., 2023) that these justifications reinforce human annotators' perception and understanding of a given task. We believe that this kind of response is helpful to non-expert annotators in improving their annotation performance.

5 Concluding Remarks

Remarkable progress in LLMs has opened up interesting possibilities in diverse domains. Accordingly, we evaluate a novel way of using LLMs as annotators. We explore the efficacy of these LLMs as a substitute or as an aid for human annotators in the context of low-resource languages, specifically Indic languages. Despite the presence of multilingual training data, including data from Indic languages data, the LLMs struggle to provide correct responses in Indic languages. We report that annotations for simpler tasks, such as sentiment analysis, can be readily recreated by the current set of LLMs. We observe that these LLMs still have a long way to go before they can be used as annotators in LR language tasks where linguistic understanding and reasoning are essential, e.g., natural language inferencing and news classification. Even though recent works have documented the feasibility of enabling annotations using these models in a positive light, these works are focused on high-resource languages. With this work, we wanted to highlight that additional efforts are needed for similar undertaking in low-resource Indic languages. In the future, we intend to employ advanced prompting strategies to aid annotations, such as using linguistic markers as knowledge prompts and in-context learning to guide the evaluations. We also hope to use back-translation to aid LLMs' understanding. We intend to experiment with these LLM annotations as weak labels to assist improvements to data collection exercises for low-resource languages. We also plan to explore the possibility of using LLMs as evaluators for quality metrics such as relevance, coherence, and fluency in the future. Furthermore, we note that the justifications provided by *ChatGPT*, along with answers, are helpful and can be further exploited for annotators' training. We plan to use these justifications to improve the prompt guidelines for LLM annotations.

Limitations We evaluate the performance of LLMs as annotators for certain tasks. There are a few limitations to note: 1) LLM performance heavily depends on the prompts. Currently, we use heuristically identified prompts, but exploring better prompts may give even better annotations in the future. 2) We agree that the experiments need more rigor. Due to restrictions on API usage, we use only a subset of available datasets. 3) We believe that quality data is also an area of concern. We use translated data in some cases, which may adversely affect the performance.

Ethics-Impact Statement

All the datasets and pre-trained models used in this work are publicly available for research purposes. The authors foresee no ethical concerns or copyright violations with the work presented in this paper.

References

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

JC Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages. *arXiv preprint arXiv:2005.02068*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: 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.

Taja Kuzman, Igor Mozetic, and Nikola Ljubešic. 2023. Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification. *ArXiv, abs/2303.03953*.

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. *arXiv preprint arXiv:2304.05613*.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.