Generative Models For Indic Languages: Evaluating Content Generation Capabilities

Savita Bhat
TCS Research
IIIT Hyderabad
savita.bhat@tcs.com

Vasudeva Varma
IIIT Hyderabad
vv@iiit.ac.in

Niranjan Pedanekar
Sony Research India
niranjan.pedanekar
@gmail.com

Abstract

Large language models (LLMs) and generative AI have emerged as the most important areas in the field of natural language processing (NLP). LLMs are considered to be a key component in several NLP tasks, such as summarization, question-answering, sentiment classification, and translation. Newer LLMs, such as ChatGPT, BLOOMZ, and several such variants, are known to train on multilingual training data and hence are expected to process and generate text in multiple languages. Considering the widespread use of LLMs, evaluating their efficacy in multilingual settings is imperative. In this work, we evaluate the newest generative models (ChatGPT, mT0, and BLOOMZ) in the context of Indic languages. Specifically, we consider natural language generation (NLG) applications such as summarization and question-answering in monolingual and cross-lingual settings. We observe that current generative models have limited capability for generating text in Indic languages in a zero-shot setting. In contrast, generative models perform consistently better on manual quality-based evaluation in Indic languages and English language generation. Considering limited generation performance, we argue that these LLMs are not intended to use in zero-shot fashion in downstream applications.

1 Introduction

Since the release of instruction-based ChatGPT, large language models (LLM) have taken the language generation research landscape by storm. Recent transformations in natural language processing (NLP) are largely enabled by pretrained LLMs such as T5 (Raffel et al., 2020), GPT3 (Brown et al., 2020), and LLaMa (Touvron et al., 2023) to name a few[1]. These models demonstrate impressive results in various NLP tasks, including language generation (NLG). Accordingly, the use of such off-the-shelf LLMs in solving downstream applications, such as conversational agents and creative copywriting, is rising. Secondly, although the performance has reached a near-human level, most of these works focus on European languages. Specifically, the latest generative models, such as ChatGPT and Bard, generate near-perfect content in English and other high-resource languages. However, English is not the native language for most of the world's population. One prime example of this is India, where people interact in one of their native languages daily. Considering India is the most populated country in the world[2], it is imperative to evaluate the latest progress in NLP (and NLG) and the potential of LLMs to be used as-is in the downstream tasks with a focus on Indic languages.

So far, the LLMs have shown considerable prowess in tackling monolingual applications. But with increasing globalization and demand for information, research in cross-lingual approaches is gaining attention. This upcoming field consists of methods to enable information access across multiple languages. With this work, we provide an initial performance evaluation in monolingual and cross-lingual settings for Indic languages.

There has been a spurt of research in the direction of evaluating generative models. Recent works include LLM evaluation in multilingual learning (Lai et al., 2023), cross-lingual summarization (Wang et al., 2023a), and multi-task, multimodal, and multilingual setting (Bang et al., 2023). Evaluating LLMs as an alternative to human annotators and evaluators is also explored in (Wang et al., 2023b; Huang et al., 2023; Törnberg, 2023; Guo et al., 2023). As a part of our analysis, we report preliminary observations on evaluating and anno-

[1] Henceforth, we use generative models and LLMs interchangeably.

[2] https://tinyurl.com/2tz9d3u2; Last accessed: 08/11/2023

Proceedings of Recent Advances in Natural Language Processing, pages 187–195
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_021

tating powers of generative models.

This work focuses on NLG tasks such as summarization and question-answering. We use LLMs such as ChatGPT variant (GPT-3.5), mT0, and BLOOMZ to evaluate zero-shot (monolingual and cross-lingual) settings. We compare the zero-shot performance of these LLMs with state-of-the-art (SOTA) baselines for the above tasks. We manually evaluate the results on quality metrics such as relevance, correctness, and fluency. We present our observations on the various generative models and generation tasks, such as summarization and question-answering in monolingual and cross-lingual settings.

The main findings of this work are as follows:

1. To the best of our knowledge, this work is the first to explore the zero-shot performance of LLMs for Indic languages. We also experiment with cross-lingual settings to analyze the correlation with the English language.

2. We observe that in terms of the ROUGE metric, the current open-source LLMs display limited performance in text generation in Indic languages. Results for cross-lingual generation (generation from Indic languages to English) also show a similar trend.

3. It should be noted that using off-the-shelf LLMs in downstream applications in Indic languages is not advisable. Results show that fine-tuned models perform far better than the zero-shot LLMs. Fine-tuning using task-specific and language-specific data is essential for better performance.

4. Content generated by LLMs is observed to be more relevant, fluent, and correct than human-generated content in mono-lingual and cross-lingual settings. It is worth noting that the manual evaluation for quality metrics reports observations contradictory to the ROUGE metric evaluation, reiterating the fact that automatic metrics do not correlate well with human evaluations.

2 Related Work

With rapid advancements in generative models, there has been a lot of interest in understanding and evaluating the performance of these models. Since many of these models have not completely disclosed their technical and data specifications (e.g., Bard and ChatGPT), experimenting in different settings is one way to test their behavior. Recently, targeted efforts have been observed to evaluate the performance of these LLMs in the context of multiple languages, modalities, and tasks. Lai et al. (2023) perform a thorough evaluation of ChatGPT for its performance in multiple languages across multiple tasks. Similarly, Bang et al. (2023) extensively investigate ChatGPT in multilingual, multimodal, and multitask setting with a focus on reasoning and hallucination. Liu et al. (2023) documents experiments evaluating ChatGPT's Text-to-SQL performance to explore its capability of generating structured SQL text for given natural language text. Wang et al. (2023a) document the performance of ChatGPT-like LLMs for cross-lingual summarization. They consider *English* and *Chinese* languages as a part of their study. In contrast, we focus solely on NLG tasks for Indic languages. We consider *English* as a part of the cross-lingual generation setup.

Using generative models for annotation or evaluation is an interesting application, and many works have been reported to explore the same. Wang et al. (2023b) explore the possibility of using ChatGPT to evaluate the quality of natural language. Guo et al.(2023) extensively investigate ChatGPT for its closeness to human experts. On similar lines, Tornberg et al. (2023) reports that ChatGPT outperforms experts and crowd-workers in annotating for certain tasks. These works consider high-resource languages such as English and Chinese. Several other works, such as (Zhu et al., 2023; Kuzman et al., 2023; Chen et al., 2023), explore using generative models as an alternative to human annotators and evaluators. Our work focuses on low-resource Indic languages to evaluate generative models for their ability to annotate and evaluate linguistic content.

3 Methodology

This work aims to evaluate the performance of generative models for NLG tasks in Indic languages in mono-lingual and cross-lingual settings. By definition, a monolingual setup considers a single language for input and output, whereas, in a cross-lingual setting, input and output content are in different languages. For example, generating an English summary from an English article is a monolingual task, while generating a Tamil summary from an English article or vice versa is a cross-lingual task. Considering continuing developments in generative models, we restrict this work to popular LLMs and selective tasks. But we are cognizant of the fact that continuous effort is required for exhaustive experimentation. In this work, we consider two broad NLG areas, viz. Summarization (SUM) and Question-Answering (QA). We use the **IndicNLG** benchmark dataset (Kumar et al., 2022) covering 11 Indic languages. These languages belong to Indo-Aryan and Dravidian language families, the main difference between them being the agglutinative nature of Dravidian languages. We also manually evaluate the generated content for quality metrics such as relevance, fluency, and correctness.

3.1 Summarization

Summarization is the process of compressing given textual content into concise and short form by preserving the most important content. It is achieved by paraphrasing or rewriting the salient information from the given input document. Recent improvements in LLMs have illustrated high-level language understanding, reasoning abilities, and fluent generation skills essential for summarization. We choose **Headline Generation** task to evaluate LLMs for their summarization capabilities. This task aims to generate a crisp and short one-sentence summary/title for a given news article.

3.2 Question-Answering

Question-Answering (QA) is a popular research area with many applications in search, recommender systems, and smart-assistants. QA systems provide a way to retrieve relevant information by querying structured and unstructured data sources. Given a user's requirements, these systems must scan given data sources, understand the query and context, collate relevant information, and apply reasoning abilities to generate appropriate responses. We seek to assess recent LLMs for their QA abilities, which will help us understand their comprehension and reasoning abilities. To this extent, we consider the following two themes for our experiments:

- *Question Generation*: generating an appropriate question for an answer and a given text content.
- *Answer Generation*: extracting an appropriate answer to a question from a given text content

3.3 Large Language Models

We explore the following LLMs in the context of Indian languages in monolingual and cross-lingual settings.

- **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 the ChatGPT platform between 11th May to 15th May 2023.
- **BLOOMZ** (Muennighoff et al., 2023) is an open-source multilingual LLM. Multitask prompted finetuning (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., 2023) 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 xP3 and xP3MT, consisting of 13 training tasks in 46 languages. Dataset xP3 uses English prompts, whereas xP3MT uses prompts that are machine-translated from English in 20 languages.

3.4 Prompting Strategies

Recent developments in generative models predominantly focus on instruction tuning with prompt engineering as the most viable method to interact with these LLMs. We heuristically design the prompting strategies for various tasks. We experimented with multiple variations of prompts, considering different paraphrases and instruction sequences. The selected prompts are chosen considering the best possible responses across different LLMs.

Summarization We consider monolingual and cross-lingual summarization for our experiments. In the following prompts, language is specified at {`lang`} and the prompts are followed by the textual content in place of {`content`} in one of the Indic languages.

- **MSUMM**:- This prompt guides LLMs to generate the summary in the same language as that of the given content:
  `I want you to act as a summarizer. I will provide the article in {lang}, and I want you`

```
to generate a one-line summary for the given article. I want the generated summary in {lang}.
Content: {content}
```

- **XSUMM**:- This prompt is used for cross-lingual summarization where content is given in one of the 11 Indic languages, and LLMs are instructed to generate a summary in English. We use the modified MSUMM prompt by changing the second {lang} to English.

Question-Answering Question-Answering task is further categorized into *Question Generation* and *Answer Generation*. We interchange the question and answer requirements according to the task. The language is specified at {lang}, context at {context} and Answer(/Question) {answer/question}.

- **MQG/MAG:**- This prompt guides LLMs to generate relevant question(answer) in the same language as that of the given context and answer(question):

```
I want you to act as a question(answer) generator. I will provide the text as a context in {lang} and an answer(question) based on the text in {lang}. I want you to generate a question(answer) for the given answer using given text as context. I want the generated question(answer) in the same language as that of the given answer(question).
Context: {context}
Answer(/Question): {answer/question}
```

- **XQG/XAG**:- This prompt is used for cross-lingual question generation and answer generation. The context is given in one of the 11 Indic languages, and LLMs are instructed to generate question (answer) in English. As earlier, we use a modified MQG/MAG prompt by using 'English' in place of the second {lang}.

Language	**Task**	
	HG	**QG**
Assamese (as)	59,031	98,027
Bengali (bn)	142,731	98,027
Gujarati (gu)	262,457	98,027
Hindi (hi)	297,284	98,027
Kannada (kn)	155,057	98,027
Malayalam (ml)	20,966	98,027
Marathi (mr)	142,590	98,027
Odia (or)	72,846	98,027
Punjabi (pa)	60,635	98,027
Tamil (ta)	75,954	98,027
Telugu (te)	26,717	98,027

Table 1: IndicNLG Benchmark datasets statistics for Headline Generation (HG) and Question Generation (QG) for 11 languages.

3.5 Quality Metrics

With the increase in popularity of NLG systems, there is a need for devising a proper way of evaluating generated content and thereby comparing systems' performances. Till now, automatic metrics such as BLEU and ROUGE are widely used even though they show little correlation with human judgment (Sai et al., 2022). In this study, we consider a randomly selected subset of articles from the *Summarization* dataset and manually evaluate the generated summaries on quality metrics such as *fluency*, *relevance*, and *correctness*. We define these metrics as follows:

Fluency refers to the correctness of the generated text with respect to grammar and word choice, including spelling (Sai et al., 2022).

Relevance evaluates whether the generated content is related to the given input data.

Correctness assesses whether the information provided in the generated content is consistent with the source or input data.

4 Experimental Setup

Datasets As mentioned earlier, we primarily use task-specific datasets from **IndicNLG** benchmark (Kumar et al., 2022). Table 1 presents data distribution for both **Headline Generation** and **Question Generation**, benchmark datasets.

To evaluate *Summarization* performance, we choose **Headline Generation** benchmark dataset from **IndicNLG** (Kumar et al., 2022). This dataset consists of news articles and corresponding headlines in Indic languages, with a total of 1,316,268 samples distributed across 11 Indic languages. We randomly select 50 samples from every 11 languages for the monolingual summarization task. To evaluate cross-lingual capabilities, we generate the output summary in English and compare it with the English translation of the corresponding reference summary.

We consider the **Question Generation** benchmark dataset from **IndicNLG** to evaluate QA capabilities. This dataset is repurposed from SQuAD question-answering dataset (Rajpurkar et al., 2016). The question, corresponding answer, and the sentence containing the answer is extracted and translated into Indic languages. The dataset consists of around 98K samples each for 11 languages. For monolingual *Question Generation* and *Answer Generation*, we randomly select 50 samples from each language-specific data samples. For cross-lingual experiments, we generate the questions in English and compare them with English translations of reference questions. Following a similar translation methodology for cross-lingual experiments, we use LLMs to generate answers in English for comparison with English translations of answers in the ground truth.

Baselines We compare the performance of zero-shot LLMs with fine-tuned **IndicBART** and **mT5** models. **IndicBART**(Dabre et al., 2022) is a pre-trained model that focuses on all 11 Indic languages considered in this work. Similarly, **mT5** (Xue et al., 2021) is a pre-trained multilingual model covering 101 languages, including Indic languages, in focus for this work. We consider results reported in (Kumar et al., 2022) for comparative analysis.

Metric Since baseline results are reported in ROUGE metric, we ROUGE-1/2/L (Lin, 2004) for our evaluation. The ROUGE score considers the lexical overlap between generated content and given reference text based on unigram (R-1), bigram (R-2), and the longest common subsequence (R-L). For ROUGE score computation, we use *multi-lingual rouge* toolkit[3].

Implementation We use official API with default settings for ChatGPT (**GPT-3.5**). For **mT0** and **BLOOMZ**, we use Huggingface checkpoints, mt0-large and bloomz-1b1, respectively. We use a subset of around 50 samples from the summarization dataset and score the corresponding generation results on the above quality metrics.

LN	GPT-3.5	mT0	BLOOMZ	mT5	IB
as	10.79	11.63	8.42	30.85	71.56
bn	11.89	7.68	8.41	31.54	39.17
gu	17.87	14.63	13.54	31.04	33.03
hi	21.22	19.68	21.11	32.55	34.57
kn	16.96	18.00	9.17	66.67	72.35
ml	13.19	13.19	12.36	39.59	60.63
mr	13.11	12.86	14.94	32.88	41.58
or	10	6.89	5.03	21.22	21.95
pa	18.64	17.69	16.11	40.13	43.81
ta	23.8	13.58	17.92	46.42	46.87
te	12.23	11.18	11.36	31.56	42.89

Table 2: Experimental results (ROUGE-L scores) for **Monolingual** *Summarization* for 11 Indic Languages (**LN**). IndicBART (**IB**) and **mT5** are finetuned state-of-the-art results.

5 Results & Analysis

In this section, we present results and analysis for *Summarization*, *Question Generation*, and *Answer Generation* tasks.

5.1 Monolingual Generation

Table 2 reports the experimental results for *summarization*, whereas Table 3 lists the results for *Question Generation*. Table 4 documents *Answer Generation* results.

Fine-tuning helps in certain tasks It can be seen that fine-tuning is extremely effective in the case of *Summarization*. We can see that the fine-tuned models, mT5 and IB, consistently show stronger performance than the zero-shot generative models in all 11 languages. In the case of *Question Generation*, the performance gap between fine-tuned models and zero-shot models is narrow, although

LN	GPT-3.5	mT0	BLOOMZ	mT5	IB
as	7.03	9.41	4.63	19.69	20.21
bn	14.6	15.36	6.94	29.56	24.49
gu	11.2	10.94	5.26	26.31	26.25
hi	22.89	22.38	11.55	34.58	32.24
kn	15.99	12.77	5.71	23.32	22.40
ml	7.34	11.99	5.08	21.82	19.71
mr	11.15	13.06	5.78	22.81	20.61
or	8.6	9.95	5.49	20.34	24.29
pa	16.11	19.64	8.95	29.72	30.59
ta	8.7	9.97	4.41	22.84	21.24
te	8.56	14.03	6.77	25.63	24.46

Table 3: Experimental results (ROUGE-L scores) for **Monolingual** *Question Generation* for 11 Indic Languages (**LN**). IndicBART (**IB**) and **mT5** are finetuned state-of-the-art results.

[3] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring; Last accessed: 08/11/2023

the fine-tuned models have better performance.

Does model architecture play a role in the performance? GPT-3.5 and BLOOMZ are decoder-only architectures, whereas mT0 is based on the encoder-decoder architecture of transformers. Although BLOOMZ performance is consistently worse in *Question Generation* and *Answer Generation*, there is no clear winner. Hence, no definite conclusion can be drawn from the observed performance results. Possible directions to evaluate are the training data size, training data sources, and prompting strategies. We keep this study for the future.

Monolingual Answer Generation is easily adaptable We observe from Table 4 that both GPT-3.5 and mT0 display strong ability to adapt to modified tasks like *Answer Generation* with comparable results. In contrast, BLOOMZ consistently lags behind, reiterating the need to analyze different LLMs in depth.

LN	GPT-3.5	mT0	BLOOMZ
as	11.88	18.91	4.79
bn	15.78	24.00	4.72
gu	14.53	24.09	4.68
hi	18.10	19.01	6.29
kn	16.91	18.88	4.59
ml	22.17	17.88	5.53
mr	15.39	20.81	5.35
or	13.21	11.36	3.78
pa	19.92	26.98	6.87
ta	16.20	21.42	6.50
te	18.04	4.97	4.47

Table 4: Experimental results (ROUGE-L scores) for **Monolingual** *Answer Generation* for 11 Indic Languages (**LN**).

5.2 Cross-lingual Generation

In cross-lingual generation experiments, we aim to generate English content corresponding to the input given in one of the Indic languages. Table 4 reports the results for the three tasks in this setting.

Adapting to cross-lingual setting. We observe that cross-lingual generation is not easily achievable using off-the-shelf generative models. Only GPT-3.5 demonstrates a strong capability for cross-lingual generation. mT0 performs equally well in *Question Generation* but lags behind in cross-lingual *Summarization* and *Answer Generation*. BLOOMZ does not adapt at all to the cross-lingual setting. One possible reason is that the generative models are unaware of such an application since it is not a part of their pre-training. Cross-lingual generation is not a typical NLG task, and hence zero-shot generative models fail to adapt for the same. We believe that additional efforts in terms of dataset and fine-tuning are necessary for better cross-lingual capabilities. Another possibility is that the prompts used in the experiments may be better suited for GPT-3.5 than the other two. We believe that more experiments with prompt engineering may improve the performance of mT0 and BLOOMZ.

5.3 Evaluation on Quality Metrics

Despite the popularity of automatic metrics like ROUGE, it is well-known that these metrics do not correlate well with human judgment for generated content quality. Figure 1 depicts the quality evaluation of generated content and corresponding average scores on each quality metric.

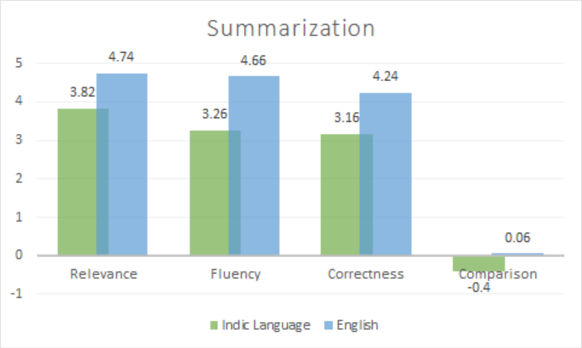

Figure 1: Manual evaluation of GPT-3.5 responses on quality metrics.

LLMs can parse the Indic languages but have better articulation in English In all quality measures, the English language generation consistently ranks higher than the generation in the Indic languages. In other words, the generative models or LLMs possess some parsing capabilities towards Indic languages, but it is not reflected in the generation process.

Generative models are better writers than humans We also compare the generated content with the reference text, with the last column representing the comparison. It can be seen that the mono-lingual generation is of lower quality as compared to the reference text, whereas the English generation is slightly better. We are conscious of the

LN	Summarization			Question-Generation			Answer-Generation		
	GPT-3.5	**mT0**	**BLOOMZ**	**GPT-3.5**	**mT0**	**BLOOMZ**	**GPT-3.5**	**mT0**	**BLOOMZ**
as	13.86	4.57	0	20.72	26.80	4.40	7.09	1.49	1.06
bn	14.11	5.17	0	23.72	29.06	4.75	10.8	1.0	1.19
gu	15.52	9.18	0.37	19.13	26.24	3.84	8.45	4.68	1.45
hi	17.49	3.41	0.62	25.25	28.933	3.90	9.46	2.97	1.35
kn	13.09	2.64	0	21.77	25.51	4.75	10.42	2.59	1.79
ml	12.3	7.98	0.19	23.99	26.12	4.22	11.01	1.91	1.52
mr	13.19	4.00	0.19	20.34	22.18	4.08	11.14	5.83	1.19
or	7.53	0	0	19.6	19.08	3.81	9.63	1.99	1.05
pa	15.4	6.00	0.006	20.95	30.06	3.51	9.87	1.64	1.33
ta	11.32	8.37	0	20.53	27.97	4.57	9.53	4.78	1.87
te	11.37	4.97	0	22	25.67	4.67	9.28	1.99	1.84

Table 5: Experimental results (ROUGE-L scores) for **Cross-lingual** *Summarization*, *Question Generation*, and *Answer Generation* for 11 Indic Languages (**LN**).

fact that extensive experiment with a large dataset is essential to establish the above observations.

5.4 Language-specific Evaluation

Aside from Hindi, all other Indic languages are categorized under low-resource or extremely-low-resource languages (Lai et al., 2023). These languages have a lower representation in the data corpus used for training LLMs. Despite that, LLMs perform comparatively well on these languages. In some cases, performance for Punjabi (pa) and Odia (or) languages is surprisingly better than that of the relatively high-resource Hindi language.

6 Concluding Remarks

With recent remarkable progress in generative models, it is essential to see no one is left behind. Advancements in low-resource languages, such as Indic languages, are lagging due to the shortage of quality data sources and technological thrust. Understanding and evaluating current progress for such under-represented languages is extremely important to identify gaps for future research. With this work, we hope to assess the generative capabilities of recent generative models in the context of Indic languages. We note that the generative models have limited capability in Indic languages in their zero-shot setting. In contrast, these models are known to perform relatively well in generating relevant English QA content highlighting their superior understanding and reasoning abilities. Off-the-shelf use of these LLMs in a zero-shot manner is observed to be suboptimal, underscoring the need for fine-tuning and task-relevant data sources. In comparison with a human evaluation of quality metrics, these models perform far better than actual reference content. It is observed that generative models may be useful as an alternative to manual annotation and evaluation efforts. We plan to continue this evaluation work by including GPT-4 and Bard. We also hope to compile more human evaluations to better understand the efficacy of generative models as an annotator or an evaluator.

7 Ethics-Impact Statement

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

Limitations We evaluate the performance of LLMs on generative tasks such as *summarization*, *question generation*, and *answer generation*. There are some limitations to note: 1) Prompts are crucial in guiding LLMs for a specific task. We have heuristically identified certain prompts, but future work could involve exploring better prompts to get better generation results. 2) We note that the evaluation comparisons need more rigor with more samples and human evaluations. Due to limitations on API usage, this work considers a subset of the dataset for comparison.

References

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*.

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.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

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*, WWW '23 Companion, page 294–297, New York, NY, USA. Association for Computing Machinery.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M Khapra, and Pratyush Kumar. 2022. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394.

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*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S Yu. 2023. A comprehensive evaluation of chatgpt's zero-shot text-to-sql capability. *arXiv preprint arXiv:2303.13547*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. pages 15991–16111.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

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*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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*.