Bengali ChartSumm: A Benchmark Dataset and Study on Feasibility of Large Language Models on Bengali Chart to Text Summarization

Nahida Akter Tanjila, Afrin Sultana Poushi, Sazid Abdullah Farhan,
Abu Raihan Mostofa Kamal, Md. Azam Hossain, Md. Hamjajul Ashmafee
NDAG Research Lab, Department of CSE,
Islamic University of Technology, Gazipur, Bangladesh
{nahidaakter, afrinsultana, sazidabdullah, raihan.kamal, azam, ashmafee}@iut-dhaka.edu

Abstract

In todays data-driven world, effectively organizing and presenting data is challenging, particularly for non-experts. Although tables organize data systematically, they often fall short in conveying intuitive insights; in contrast, charts provide clear and compelling visual summaries. Although recent advances in NLP, powered by large language models (LLMs), have primarily benefited high-resource languages such as English, low-resource languages such as Bengali spoken by millions worldwide still face significant data limitations. This research addresses this gap by introducing "Bengali ChartSumm," a benchmark dataset with 4,100 Bengali chart images, metadata, and summaries. This dataset facilitates the analysis of LLMs (mT5, BanglaT5, Gemma) in Bengali chart-to-text summarization, offering essential baselines and evaluations that enhance NLP research for low-resource languages.

1 Introduction

In today's data-driven world, enormous amounts of data are generated every second, presenting unique challenges in organizing and presenting it effectively. Although tabular formats can organize data, they are often inadequate for complex datasets, particularly for non-experts who may find it challenging to identify essential insights. This difficulty arises because tables lack intuitive trends or highlights, making it challenging to extract valuable information. To address this, various tools and methods have been developed to uncover hidden patterns in data efficiently. Among these, specialized visualizations, particularly charts, stand out as powerful tools for translating complex information. By blending numerical data, text, and visual elements, these visualizations communicate intricate information in a clear and accessible way, making data insights more understandable and impactful (Islam et al., 2021).

The emergence of deep learning-driven artificial neural networks has significantly enhanced Natural Language Processing (NLP), boosting the efficiency and accuracy of textual data processing. However, this progress has been largely limited to high-resource languages like English, which benefit from vast amounts of labeled and unlabeled data from sources such as books, social media, websites, and academic publications. This data-intensive environment allows NLP models to be more optimized for specific tasks, resulting in improved accuracy. In contrast, low-resource languages lack this support, making it challenging to develop the NLP domain due to the absence of even baseline datasets, models, and evaluation benchmarks. Essential resources like tokenizers, parsers, part-of-speech taggers, and dependency grammars are often missing or underdeveloped for these languages, and creating them from scratch requires substantial linguistic expertise and time.

Bengali is one of the most widely spoken languages in the world, with approximately one in eight people worldwide using it but research resources for this language are still scarce in comparison to its substantial population (Ekram et al., 2022). Although English has made significant advances in the field of natural language processing (NLP) in various chart-related downstream tasks such as question answering, summarization, and the mathematical analysis of chart images, Bengali and other low-resource languages are facing severe data limitations. Although Bengali speakers frequently employ Bengali charts in a variety of contexts, these constraints impede the efficient training and fine-tuning of NLP models. Text summaries, in particular, can enhance the comprehension and interpretation of charts by highlighting key elements like temporal trends, causal relationships, and evaluative aspects (Bhattacharjee et al., 2023). Given the abundant resources available in English, there is an opportunity to leverage them to address this

Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), pages 35–45
January 19, 2025.

research gap and create data repositories that will allow domain experts to access and utilize chart data more effectively in the Bengali language.

Large Language Models (LLMs) have made remarkable progress in natural language processing (NLP), particularly in language generation and other language-centric tasks. Multimodal LLMs, specifically vision-language models trained on chart data, excel at tackling the challenges of integrating visual and textual information, making them highly effective in tasks that require comprehensive understanding across modalities. These vision-language models undergo extensive pre-training on chart-related tasks, involving multitask instructional tuning and task-specific fine-tuning, which equips them to perform well across a range of downstream tasks.

However, most of the research in this field focuses on high-resource languages such as English, leaving low-resource languages like Bengali under-represented (Kabir et al., 2023). This lack of research and the absence of baseline models and evaluation benchmarks make it challenging to develop effective NLP models for these languages. Recently, multilingual models have been introduced to address these limitations, but they still suffer from the under-representation of low-resource languages, highlighting a significant opportunity to explore and establish robust baseline research for these languages.

To the best of our knowledge, large language models (LLMs) have not yet been specifically applied to the task of chart-to-text summarization in Bengali. This is primarily due to the lack of extensive, well-defined datasets that include chart images, metadata, and detailed summaries in Bengali. In this work, we have addressed the scarcity of public datasets in the automatic chart summarization task. We have proposed a benchmark dataset - Bengali ChartSumm comprising 4,100 chart images with corresponding chart metadata and summaries and conducted a study on large language models' feasibility on Bengali Chart Summarization. A sample from the curated dataset is shown in Figure: 1 as an example.

2 Literature Review

The task of generating descriptive summaries from non-linguistic structured data, such as tables or charts, is referred to as "chart-to-text summarization." This falls within the broader domain of natural language generation (NLG) and involves converting data visualizations (like line, bar, bubble, and pie charts) into textual descriptions. Current chart-to-text summarization systems produce summaries based either on the chart image itself (Hsu et al., 2021) or on the metadata associated with the chart (Obeid and Hoque, 2020; Kantharaj et al., 2022).

Several datasets have been developed to support research in English chart to text summarization, including SciCap (Tan et al., 2022), Chart2Text (Obeid and Hoque, 2020), AutoChart (Zhu et al., 2021), and Chart-To-Text (Kantharaj et al., 2022). These datasets vary in their formulation, ranging from table to text descriptions, image to extracted metadata using OCR to text descriptions, and system-generated short summaries to descriptive human-written summaries. The development of these datasets reflects the growing interest in exploring automatic chart summarization techniques and evaluating their performance across different types of charts and data sources (Rahman et al., 2023).

Deep learning-based techniques have recently gained substantial attention (Obeid and Hoque, 2020; Zhu et al., 2021; Kantharaj et al., 2022) due to their improved performance over traditional template-based approaches. However, the lack of datasets for chart-to-text summarization poses a significant challenge: not only do models for this task need refinement, but also their overall effectiveness has yet to be fully assessed. Most multimodal foundation models (Li et al., 2023) are focused on natural images and have achieved remarkable progress in fields like image captioning (Vinyals et al., 2015) and visual question answering (Johnson et al., 2017).

Some approaches have adapted vision-language models for chart-related tasks (Han et al., 2023) or created plugins enabling large language models (LLMs) to interpret charts (Xia et al., 2023). Transfer learning, facilitated through learned language representations in models like T5 and GPT that use transformer architectures, has become integral to NLP, allowing language models to be adapted for a range of downstream tasks (summarization, question-answering, inference, etc.) (Raffel et al., 2020). Nevertheless, challenges persist for low-resource languages, which face under-representation, biases, lack of required datasets for downstream tasks, and limited evaluation benchmarks for such models (Pires et al., 2019).

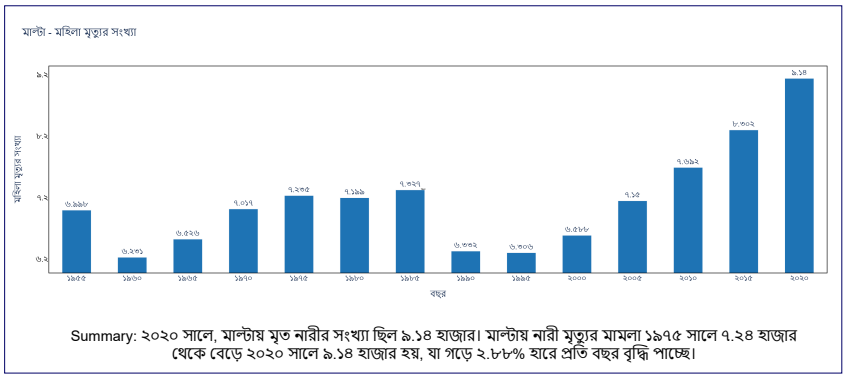

Figure 1: An example of the Bengali Chart image and corresponding summary.

3 Bengali Chart-to-text dataset

The lack of benchmark chart-to-text datasets in Bengali has compelled us to utilize language translation services for transforming existing chart-to-text datasets from resource-abundant English into Bengali, as illustrated in Figure: 2 mostly extending the work by (Rahman et al., 2023). By translating a part of their dataset to Bengali, we developed a resource tailored for Bengali language processing. Subsequently, we executed further steps to finalize the curation of this dataset, preparing it for the fine-tuning and evaluation of large language models.

3.1 Data Collection

Finding an appropriate data source is the first step in this research as it is essential to compile an extensive dataset for the low-resource Bengali language. As an available resource for summarization tasks from chart images, our first step was to utilize an English chart-to-text dataset (Rahman et al., 2023) collected from different public online repositories, including both long and short summaries. Although this repository includes bar, line, and pie charts, we chose to include only line and bar charts in our dataset for ease of curation and model fine-tuning. In our next phase, we will include pie charts and other common charts in Bengali from many more public repositories to make it diverse, realistic, and challenging for research. We started translating about 4,100 samples from the original dataset covering diverse topics while maintaining fidelity to the original content as well as ensuring linguistic and cultural accuracy. These translations included both titles and captions, indicating a substantial increase in the dataset's usefulness.

3.2 Data Annotation

Following machine translation, the annotation process was meticulously performed by human annotators (Munaf et al., 2023). We selected undergraduate students with STEM backgrounds, specifically those with strong chart analysis skills and fluency in both English and Bengali. From a group of interested candidates (19 students), we conducted a controlled assessment to evaluate one's competency in translating and analyzing charts and identify expert annotators, ultimately selecting five students based on their performance. To ensure the dataset's quality and relevance for the intended audience, random samples of their annotations were reviewed, verifying the accuracy and cultural appropriateness of translations. The use of a relatively small group of annotators from similar educational backgrounds introduces the potential for biases. To mitigate this, random quality checks were performed on their annotations to verify accuracy and cultural appropriateness. In addition, random samples of annotations were periodically reviewed to ensure that translated summaries preserved the intent and tone of the original dataset while being accessible to a Bengali-speaking audience. Some samples are shown in Appendix A.1.

3.3 Dataset Preparation

After annotating the dataset, we conducted several **data preprocessing** steps to prepare it for analysis (Meng et al., 2024). This process included cleaning the data by removing whitespace, newlines, and irrelevant content, as well as converting metadata into a format compatible with language models. Additionally, heuristic rules were applied to generate x-labels and y-labels where the originals were corrupted. In the following **tokenization**

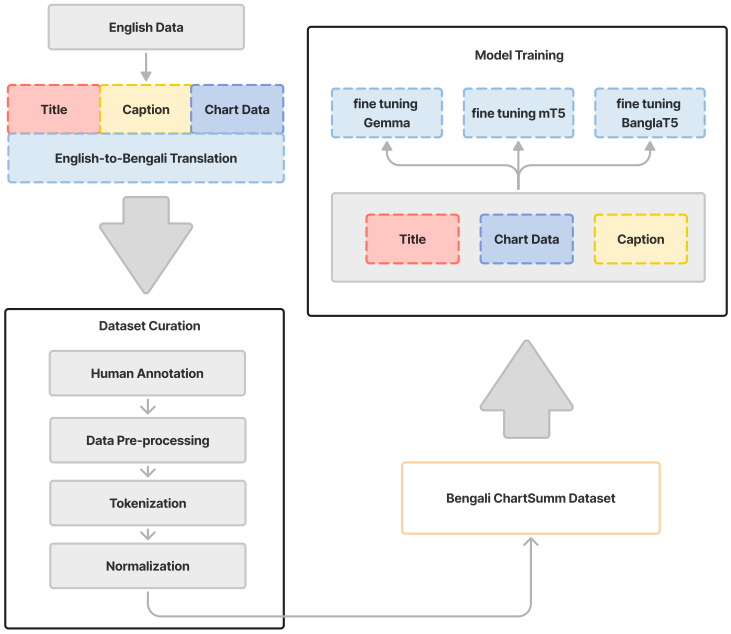

Figure 2: A comprehensive overview of this research.

LLM Models	Number of Parameters
Google/mT5-small[1]	300 Millions
BanglaT5[2]	247 Millions
Google/gemma-2b[3]	2.5 Billions[4]

Table 1: LLM models used for this experiment and corresponding number of parameters in the models.

step, we used customized tokenizers tailored to the pre-trained models (as detailed in Table 1) to stem and tokenize the text data, ensuring it was ready for subsequent analysis (Rahman et al., 2023).

Data normalization was done as Bengali words have many characters whereas English has only 26 characters. As a result, a word could have different forms having identical appearances and meanings. In Bengali, a special character "'়'" having the symbol called "Nukta" which can alter the meaning of a character depending on its inclusion or absence. Various tokenizers approach this issue differently to minimize confusion caused by multiple representations of the same word. To resolve this, we utilized a Bengali text normalizer (Hasan et al., 2020), which effectively handles these challenges, including the accurate processing of Bengali numerical entities.

3.4 Dataset Analysis

We analyzed the text length distribution which gives useful information about the dataset's features. The visualization aids in understanding the variety in text lengths among different types of content in the training data, which is critical for model tuning which is seen in Figure 3. The distribution is relatively narrow, with 50% of titles falling between 39 and 60 tokens. The histogram shows a peak in frequency of around 50 tokens, indicating that most titles are of moderate length, likely brief descriptions or names. Summaries exhibit a wider distribution, suggesting varying levels of detail in the chart descriptions. The middle 50% of summaries fall between 138 and 255 tokens. The histogram indicates a concentration of around 200 tokens, with a somewhat even spread, showing summaries are typically more detailed and descriptive than titles but less than the chart data.

In our chart dataset, text length shows the greatest variability, with an average length of approximately 586 tokens but a very high standard deviation. The histogram reveals a multi-modal distribution, with peaks around both lower and higher text lengths,

[1]https://huggingface.co/google/mt5-small
[2]https://huggingface.co/csebuetnlp/banglat5
[3]https://huggingface.co/google/gemma-2-2b-it
[4]Both embedding and non-embedding parameters

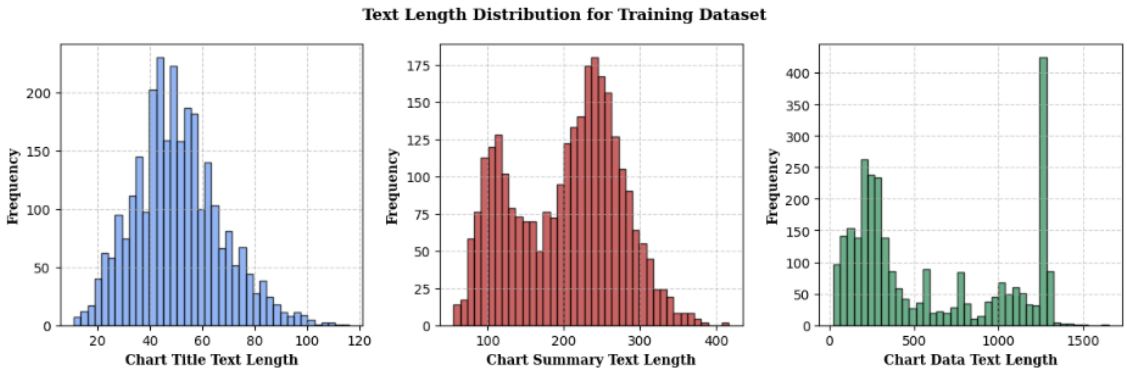

Figure 3: Visualize text length distribution for Training dataset

indicating that some charts have very concise data, while others provide extensive details, which is why we have handpicked the text with a stable and average number of tokens that will help in our model to not be over-fitting. Addressing this variability could improve the feasibility of applying LLMs to this complex task, making it an essential consideration in developing *Bengali ChartSumm* as a benchmark dataset.

4 Methodology

This section outlines the process of fine-tuning, training, and evaluating our selected models (shown in Figure: 2) for the Bengali chart-to-text summarization task. The primary focus is on the setup of the evaluation metrics and the rationale for their selection, ensuring a robust assessment framework.

4.1 Problem Formulation

The overall process of our Bengali Chart-to-Text Summarization system is shown in Figure 2. We consider the problem setting for Bengali chart summarization assuming that the underlying data table of the chart image is available. Formally, we are given a dataset with N examples $D = \{m_i, s_i\}_{i=1}^{N}$, where m_i represents the underlying metadata for a given chart image and s_i represents the target summary text for m_i. The Bengali Chart Summarization models learn to predict the summary s_i given the metadata m_i.

4.2 Model Training and Fine-tuning

Each model was fine-tuned on our prepared *Bengali ChartSumm* dataset to optimize its performance for generating Bengali text summaries from chart metadata. For this task, we utilized three models: mT5 (Xue et al., 2020), BanglaT5 (Bhattacharjee et al., 2023), and Gemma (Team et al., 2024), each with unique architectures and strengths for multilingual and Bengali text generation (listed in Table 1).

For the models mt5 and banglat5, the input format was structured as "title + chart metadata", with the output targeted as the "caption". We defined a custom PyTorch *Dataset* class, *Seq2SeqDataset*, which tokenizes and processes input data. The dataset class concatenates the chart title and chart data into a single input text sequence with the format <Title> "চার্ট তথ্য:" <Chart Data>. The summary column serves as the target sequence.

For the Gemma model, we used prompting and context to structure the dataset to get the summary as a prompt answer. We proceeded by loading the Bengali chart data, which consisted of three text files: titles, summaries, and chart data. The data was organized into a DataFrame with three columns: *title*, *summary*, and *chart_data*. We created a custom template to serve as an instruction for the model. This template was designed to instruct the model to provide chart summarization in Bengali, with a format that included the chart title, and chart data as `<input_sequence>`. The complete prompt template was like this -

```
"Instruction: Your task is to give chart
summarization in Bengali language:
(এই চার্টটি বাংলা ভাষায় সারাংশ কর):"
"Given data: চার্ট তথ্য:" <input_sequence>
"Summary: সারাংশ:"
```

Each data line was modified to add labels for the x-axis and y-axis, making it easier for the model to understand the context. A regular expression function, *modify_data_line*, was used to add these

labels to the data points, creating a standardized input format. Using this template, the chart titles, data, and corresponding captions were combined into structured prompts and stored in a list format suitable for model fine-tuning.

4.3 Evaluation Metrics

To rigorously evaluate the quality of the generated summaries, we employed a set of commonly used metrics in natural language processing for summarization and translation tasks (Rahman et al., 2023; Meng et al., 2024). These metrics assess various aspects of model performance by measuring both n-gram overlaps and error rates. **BLEU** (Bilingual Evaluation Understudy) (Post, 2018) measures n-gram precision between machine-generated and reference texts, emphasizing precision to evaluate how closely the generated summaries align with the reference summaries. **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), often used in text summarization and longer text generation tasks, focuses on recall by measuring the overlap of n-grams or sequences between machine-generated and reference texts. We report three variants of **ROUGE**: **ROUGE-1**, which assesses unigram recall; **ROUGE-2**, which evaluates bigram recall, reflecting phrase-level accuracy; and **ROUGE-L**, which captures the longest common subsequence, providing a measure of structural similarity. **CER** (Character Error Rate) (Wang et al., 2016) calculates the ratio of character-level errors (insertions, deletions, substitutions) to the total characters in the reference text, making it especially important in Bengali text generation, where small character variations can significantly alter meaning. **WER** (Word Error Rate) (Ali and Renals, 2018) measures word-level accuracy by calculating insertions, deletions, and substitutions relative to the reference text, with a lower WER indicating fewer errors in capturing the intended word sequence. Together, these metrics offer a comprehensive understanding of the models summarization quality, highlighting both precise word matching and broader structural accuracy, laying the groundwork for a thorough performance comparison of the models, which is presented in the following section.

5 Experimental Setup

To aid model training, we normalized the text input with the normalize function from the normalizer library (Hasan et al., 2020), which is specially developed for Bengali language processing. Each input sequence is tokenized with a maximum length of 256 tokens converted into PyTorch tensors. Training configurations for each model included 15 epochs, a learning rate of 1e-3, epsilon of 1e-8, weight decay of 0.01, AdamW[1] optimizer and a batch size of 16, optimized for the given dataset and task requirements. The dataset was split into a Training set: 70%, a Validation set: 10.5%, and a Test set: 19.5%. We also applied appropriate approaches for the remaining model configurations. For evaluation, we loaded the fine-tuned model and tokenizer and generated summaries for the test dataset. We used the generate function with num_beams (4) to apply beam search and set max_length (512) for generated summaries. To avoid overly short summaries, we omitted any length penalty and removed early stopping to ensure complete generation.

6 Evaluation

6.1 Results and Discussions

In this section, we analyze the performance of three distinct Large Language Models: mT5, BanglaT5, and Gemma using several key evaluation metrics. These models were tested on their ability to generate coherent and accurate text based on given reference texts. The evaluation metrics considered include ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), which measures the recall of n-grams and the longest common subsequence, BLEU (Bilingual Evaluation Understudy), which focuses on n-gram precision, and two error rates: CER (Character Error Rate) and WER (Word Error Rate)which gauge the accuracy of the generated text at the character and word levels, respectively. The summary of this evaluation is shown in Table 2.

6.1.1 ROUGE Scores

BanglaT5 outperforms mT5 and Gemma in ROUGE-1, achieving a score of 0.0678 compared to mT5's 0.0422, marking a significant 63% improvement. This suggests that BanglaT5 is better at recalling individual words (unigrams) from the reference text. However, when examining ROUGE-2 and ROUGE-L, all models exhibit lower scores, indicating struggles with recalling bigrams and maintaining the overall structure of the reference text.

[1]https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html

Models	Rouge-1 ↑	Rouge-2 ↑	Rouge-L ↑	BLEU ↑	CER ↓	WER ↓
mT5	0.0422	0.0015	0.0397	**0.2779**	**0.5295**	**0.7189**
BanglaT5	**0.0678**	**0.0016**	**0.0592**	0.0505	0.9681	1.1530
Gemma	0.0227	0.0008	0.0204	0.2153	0.7828	1.0653

Table 2: Experiment results on Large Language Models

These low scores highlight the challenges faced by the models in producing text that accurately mirrors the phrasal patterns and the longest common subsequences found in the reference material.

6.1.2 BLEU Score

In contrast to the ROUGE scores, mT5 shows a significant advantage in the BLEU score, with 0.2779 compared to BanglaT5's 0.0505 and Gemma's 0.2153. The BLEU metric emphasizes n-gram precision, suggesting that mT5 is more successful at generating text with sequences of words that exactly match those in the reference text. Despite this, the low ROUGE-L score indicates that while mT5 may produce some n-gram matches, it may still struggle with overall fluency and faithfulness to the reference text's structure.

6.1.3 CER and WER

When considering CER (Character Error Rate) and WER (Word Error Rate), BanglaT5 displays higher error rates (0.9681 for CER and 1.1530 for WER) compared to mT5 (0.5295 for CER and 0.7189 for WER) and Gemma (0.7828 for CER and 1.0653 for WER). These metrics reflect the tendency to introduce more character and word errorssuch as insertions, deletions, or substitutions when generating text, leading to a higher deviation from the reference for BanglaT5 and Gemma.

6.1.4 Overall Comparison

The comparative analysis of the models suggests a trade-off between recall and precision. BanglaT5, while better at recalling individual words, struggles with n-gram precision, fluency, and overall correctness, as evidenced by its lower BLEU score and higher error rates. On the other hand, mT5, despite its lower ROUGE-1 score, performs better in BLEU, CER, and WER, indicating fewer errors and better n-gram precision, although it may lack in producing text that fully captures the structure of the reference text. These results imply that the training data for BanglaT5 might have been less comprehensive or diverse, limiting its ability to generate text with complex phrasings and structures. While all models have their respective shortcomings, mT5 demonstrates a more balanced performance, particularly in generating summaries that are more precise and contain fewer errors. The findings underscore the importance of model training on diverse and high-quality datasets to improve the overall performance of text generation models. Appendix A.2 provides a sample summary generated by Gemma as part of a qualitative comparison.

7 Conclusion

Our study on Bengali chart-to-text summarization establishes a strong foundation for future research and development in automatic chart summarization, particularly for under-resourced languages like Bengali. The primary objectives were to create a dataset specifically designed for Bengali chart summarization and to evaluate the applicability of Large Language Models (LLMs) in this context. By addressing the resource gap and demonstrating the potential of LLMs in Bengali chart summarization, this work not only advances the field but also paves the way for further research and development. Showing the feasibility of using language models for Bengali chart-to-text summarization represents a significant step in applying these powerful models to underrepresented languages and tasks, ultimately promoting inclusivity and accessibility in NLP technology.

Future research could focus on integrating Bengali-specific Optical Character Recognition (OCR) technology to enhance the accuracy of reading both handwritten and printed Bengali text in charts, thereby expanding the scope of chart formats our dataset and models can handle. Expanding the dataset to include a broader range of common chart types (e.g., pie charts, bubble charts, etc.) could increase its diversity and complexity. Additionally, utilizing state-of-the-art large language models (LLMs), such as LLaMA and Mistral, could significantly enhance the performance of chart summarization. On the other hand, translating the dataset and models into other languages would broaden their linguistic applicability and enable

cross-lingual comparisons, paving the way for universally applicable chart summarization methods. Another promising avenue is the development of interactive platforms or chatbots that use these models to provide real-time interpretation of Bengali charts, increasing accessibility for users. Despite these advancements, further work is needed to address the technical challenges associated with OCR technology tailored for Bengali and other languages.

Limitations and Ethical Considerations

While Bangla and English include various chart types, for this research, we focused on bar and line charts due to the availability of public English datasets and the relative simplicity of these chart types. Our LLM models are trained with limited resources, which restricts their efficiency in generating complex, detailed summaries. Increasing dataset diversity and sizes, along with more intensive training on state-of-the-art multimodal LLMs, could better support complex reasoning across visual and textual features, leading to improved summarization performance.

We addressed several ethical considerations during the dataset collection and annotation stages. To respect the intellectual property rights of dataset sources, we exclusively used publicly available charts that comply with their terms and conditions. Our annotators were compensated above the minimum wage in Bangladesh (12,500 taka, or approximately $113 USD per month). Each task was estimated to take 3-5 minutes, and annotators were paid 2.5 taka ($0.021 USD) per task. Additionally, to protect their privacy, all annotations were anonymized. For reproducibility, our experimental hyperparameter settings are provided in Section 5.

Acknowledgment

This work was supported by the supported by the Islamic University of Technology Research Seed Grants (IUT RSG) (Ref: REASP/IUT-RSG/2022/OL/07/013). We extend our sincere gratitude to our annotators for their invaluable support in creating the dataset. We are also deeply thankful to Raian Rahman and Ishmam Tashdeed, alumni of the NDAG Research Lab, for their constructive feedback. Additionally, we appreciate the thoughtful comments and suggestions from the anonymous reviewers, which greatly contributed to improving this paper.

References

Ahmed Ali and Steve Renals. 2018. Word error rate estimation for speech recognition: e-WER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 20–24, Melbourne, Australia. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, and Rifat Shahriyar. 2023. Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 714–723.

Syed Mohammed Sartaj Ekram, Adham Arik Rahman, Md. Sajid Altaf, Mohammed Saidul Islam, Mehrab Mustafy Rahman, Md Mezbaur Rahman, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2022. BanglaRQA: A benchmark dataset for under-resourced Bangla language reading comprehension-based question answering with diverse question-answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2518–2532, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. Chartllama: A multimodal LLM for chart understanding and generation. *CoRR*, abs/2311.16483.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for bengali-english machine translation. *arXiv preprint arXiv:2009.09359*.

Ting-Yao Hsu, C. Lee Giles, and Ting-Hao Kenneth Huang. 2021. Scicap: Generating captions for scientific figures. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3258–3264. Association for Computational Linguistics.

Md. Rafiqul Islam, Jiaming Zhang, Md. Hamjajul Ashmafee, Imran Razzak, Jianlong Zhou, Xianzhi Wang, and Guandong Xu. 2021. Exvis: Explainable visual decision support system for risk management. In *8th International Conference on Behavioral and Social Computing, BESC 2021, Doha, Qatar, October 29-31, 2021*, pages 1–5. IEEE.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society.

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2023. Benllmeval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp. *arXiv preprint arXiv:2309.13173*.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. 2024. Chartassisstant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *CoRR*, abs/2401.02384.

Mubashir Munaf, Hammad Afzal, Naima Iltaf, and Khawir Mahmood. 2023. Low resource summarization using pre-trained language models. *CoRR*, abs/2310.02790.

Jason Obeid and Enamul Hoque. 2020. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 138–147, Dublin, Ireland. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4996–5001. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. 2023. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *arXiv preprint arXiv:2304.13620*.

Hao Tan, Chen-Tse Tsai, Yujie He, and Mohit Bansal. 2022. Scientific chart summarization: Datasets and improved text modeling. In *SDU@ AAAI*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164. IEEE Computer Society.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. Character: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510.

Renqiu Xia, Bo Zhang, Haoyang Peng, Ning Liao, Peng Ye, Botian Shi, Junchi Yan, and Yu Qiao. 2023. Structchart: Perception, structuring, reasoning for visual chart understanding. *CoRR*, abs/2309.11268.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Jiawen Zhu, Jinye Ran, Roy Ka wei Lee, Kenny Choo, and Zhi Li. 2021. Autochart: A dataset for chart-to-text generation task. *Preprint*, arXiv:2108.06897.

A Appendix

A.1 Human Annotations

Table 3 presents a qualitative comparison highlighting specific tasks performed by our human annotators to capture and preserve the nuances of the language.

English Summary	Machine Translation	Human Annotation	Comments
Between 1965 and 2014, Guinea Bissau CO2 emissions from gas flaring remained stable at around 0 thousand metric tons.	১৯৬৫ থেকে 2014 সালের মধ্যে, গিনি বিসাউ এর উদ্দীপ্ত গ্যাস থেকে CO2 নির্গমন প্রায় ০ হাজার মেট্রিক টনে স্থিতিশীল ছিল।	১৯৬৫ থেকে ২০১৪ সালের মধ্যে, গিনি বিসাউ এর উদ্দীপ্ত গ্যাস থেকে কার্বন ডাই অক্সাইড নির্গমন প্রায় ০ হাজার মেট্রিক টনে স্থিতিশীল ছিল	Replace “CO2” with its full form and the year “2014” with its proper Bengali number (in summary)
Italy - Total population aged 25-49 years	ইতালি - মোট জনসংখ্যা ২৫-৪৯ বছর বয়সী	ইতালি - ২৫-৪৯ বছর বয়সী মোট জনসংখ্যা	Rearrange words to clarify in natural Bengali (in title)
In 2019, female life expectancy for Australia was 85 years. Female life expectancy of Australia increased from 74.4 years in 1970 to 85 years in 2019 growing at an average annual rate of 0.27%.	২০১৯ সালে, অস্ট্রেলিয়ায় মহিলাদের আয়ু ছিল ৮৫ বছর। অস্ট্রেলিয়ার মহিলাদের আয়ু ১৯৭০ সালে ৭৪.৪ বছর থেকে বেড়ে ২০১৯ সালে ৮৫ বছর হয়েছে যা বার্ষিক গড় ০.২৭% হারে বৃদ্ধি পেয়েছে।	২০১৯ সালে, অস্ট্রেলিয়ায় মহিলাদের প্রত্যাশিত আয়ুষ্কাল ছিল ৮৫ বছর। অস্ট্রেলিয়ার মহিলাদের প্রত্যাশিত আয়ুষ্কাল ১৯৭০ সালে ৭৪.৪ বছর থেকে বেড়ে ২০১৯ সালে ৮৫ বছর হয়েছে যা বার্ষিক গড় ০.২৭% হারে বৃদ্ধি পেয়েছে।	Use "প্রত্যাশিত আয়ুষ্কাল" instead of "আয়ু," to translate the word "life expectancy" more precisely (in summary)
Monthly oil production in Angola 2019-2021	অ্যাঙ্গোলা ২০১৯-২০২১ এ মাসিক তেল উত্পাদন	অ্যাঙ্গোলা ২০১৯-২০২১ এ মাসিক তেল উৎপাদন	Refine Bengali spelling (in title)
The Christmas of 2020 is being perceived differently among all the people celebrating it. According to a survey conducted in Italy, 85 percent of the respondents strongly agreed or agreed that this Christmas is going to be different from former Christmases. Some 66 percent believed it is going to be sadder, while 25 percent thought it would be more simple.	২০২০ এর ক্রিসমাস এটি উদযাপন করা সমস্ত লোকের মধ্যে আলাদাভাবে বিবেচিত হচ্ছে। ইতালিতে পরিচালিত একটি সমীক্ষা অনুসারে, ৮৫ শতাংশ উত্তরদাতারা দৃ strongly ়ভাবে সম্মত বা সম্মত হয়েছেন যে এই ক্রিসমাসটি প্রাক্তন ক্রিসমাস থেকে আলাদা হতে চলেছে। প্রায় ৬৬ ৬৬ শতাংশ বিশ্বাস করেছিলেন যে এটি দুঃখজনক হতে চলেছে, যখন ২৫ শতাংশ ভেবেছিলেন এটি আরও সহজ হবে।	২০২০ এর ক্রিসমাস এটি উদযাপন করা সমস্ত লোকের মধ্যে আলাদাভাবে বিবেচিত হচ্ছে। ইতালিতে পরিচালিত একটি সমীক্ষা অনুসারে, ৮৫ শতাংশ উত্তরদাতারা দৃঢ়ভাবে সম্মত বা সম্মত হয়েছেন যে এই ক্রিসমাসটি প্রাক্তন ক্রিসমাস থেকে আলাদা হতে চলেছে। প্রায় ৬৬ শতাংশ বিশ্বাস করেছিলেন যে এটি দুঃখজনক হতে চলেছে, যখন ২৫ শতাংশ ভেবেছিলেন এটি আরও সহজ হবে।	Refine to preserve the intent (in summary)

Table 3: Tasks done by human annotators.

A.2 Chart Summarization by LLM

Figure 4 presents a sample summary generated from the provided chart data and a prompt by Gemma, alongside its corresponding gold label. This demonstrates that straightforward trends are readily identified by the models and effectively represented in the Bengali summary.

Instruction: Your task is to give chart summarizaton in Bengali language (এই চার্টটি বাংলা ভাষায় সারাংশ কর):

চার্ট তথ্য: জর্জিয়া - সম্পূর্ণ নিয়ন্ত্রণে সেচের জন্য সজ্জিত এলাকা x-axis: বছর y-axis: সম্পূর্ণ_নিয়ন্ত্রণে_সেচের_জন্য_সজ্জিত_এলাকা chart_data: ১৯৯৬.০ ৪৩৭.৫ ১৯৯৭.০ ৪৩৮.১ ২০০২.০ ৪৪১.০ ২০০৭.০ ৪৭২.২ ২০১২.০ ৪৭২.২ ২০১৭.০ ৪৭২.২ ২০১৮.০ ৪৭২.২

সারাংশ: ৪৭২.২ বর্গ কিলোমিটার জমিতে ১৯৯৬ থেকে ২০১৮ সাল পর্যন্ত সম্পূর্ণ নিয়ন্ত্রণে সেচের জন্য সজ্জিত এলাকা ছিল জর্জিয়াতে।

Gold label: ২০১৮ সালে সম্পূর্ণ নিয়ন্ত্রণে সেচের জন্য সজ্জিত জর্জিয়া এলাকা ৪৭২.২ কিলোমিটার স্তরে ছিল, যা আগের বছরের থেকে অপরিবর্তিত ।

Figure 4: A Bengali chart summary generated by Gemma.