What Matters for LLMs in Low-Resource Language? A Survey of Bangla Language Processing Datasets, Adaptation Techniques, and Application Areas

Anonymous ACL submission

Abstract

Large Language Models (LLMs) show tremendous abilities in every field. To fully assess the usefulness of LLMs, it is essential to evaluate their performance in low-resource languages and ensure that improvements in accuracy do not come at the cost of increased bias or hallucination. We focus on Bangla language because, despite being the world's 7th most spoken language, there remains a lack of comprehensive reviews on this topic. Focusing on a single language also allows us to more deeply investigate *what actually matters for making LLMs work well in Bangla?* We systematically analyze recent work on Bangla language processing with LLMs, reviewing benchmarks, datasets, monolingual and multilingual model families, adaptation strategies, and application areas. We conclude by presenting clear research directions that not only guide Bangla researchers but also have implications for other low-resource and morphologically rich languages like Bangla.

1 Introduction

Imagine asking a smart assistant such as ChatGPT a math reasoning question in English at a 8th-grade level. It would likely provide a detailed and accurate answer. However, if you ask the same question in a low-resource language like Bangla, also known as Bengali, which is spoken by nearly 300 million people worldwide, the response might be incorrect, confusing or lacking clarity, as shown in Figure 1. This gap is not limited to a single math problem. Across a wide range of tasks, Bangla responses frequently degrade in quality, highlighting a broader global inequity in LLMs' performance. About 4.8 billion people (Roxas and Recario, 2024), or roughly 59 percent of the world's population, speak languages other than the few high-resource ones. Yet, LLM training data is dominated by high-resource languages (Imani et al., 2023). This imbalance creates a significant technological gap. It is limiting access to information and fair treatment for speakers of Bangla.

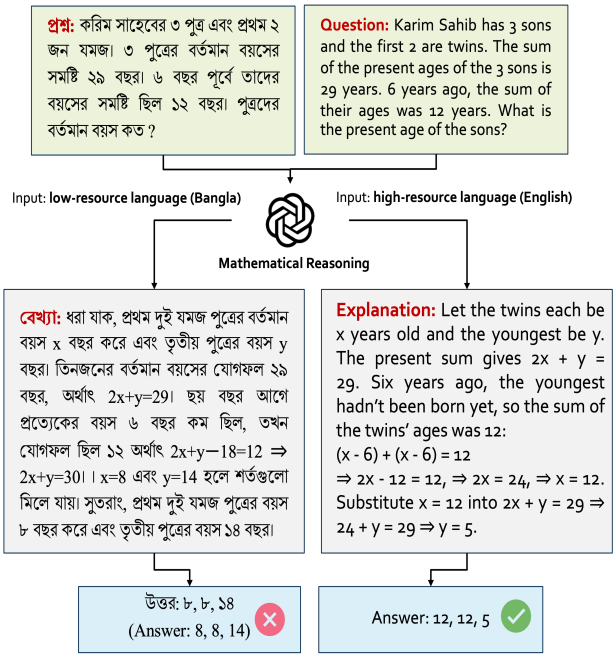

Figure 1: LLMs correctly solve mathematical reasoning problems in English but not in Bangla. The example is from BanglaMATH (Prama et al., 2025), with responses generated by ChatGPT-4o on November 4, 2025.

In today's world, LLMs have become more important than ever in history. To improve the performance of LLMs for Bangla, a systematic survey is essential. Over the years, several studies have explored Bangla language research. For instance, it Alam et al. (2021) provides a review of Bangla NLP tasks and the utility of transformer models. A comparative analysis of machine learning and deep learning methods for Bangla NLP is presented in (Sen et al., 2022). In addition, reviews have focused on specific tasks such as preprocessing techniques (Chakraborty et al., 2025), question generation (Faisal et al., 2024a), and natural language interfaces (NLI; Faria et al., 2025) using transformer models for Bangla. However,

a comprehensive survey that addresses the recent limitations and future directions of Bangla using LLMs across all aspects is still missing.

We have conducted a survey that is both timely and useful. First, this field is evolving rapidly, with new datasets, model variants and adaptation strategies for LLMs being introduced every few months. For instance, in 2025 alone, more than ten new datasets have been published (Prama et al., 2025; Fayaz et al., 2025; Mahtab et al., 2025; Islam and Sarti, 2025; Joy, 2025; Haider et al., 2025; Paul et al., 2025), and four new language models have been released (Bhattacharyya and Bhattacharya, 2025; Raihan and Zampieri, 2025; Nahin et al., 2025; Zehady et al., 2024). Second, existing results are scattered and difficult for researchers to compare. Third, even small design decisions, such as updating the tokenizer or applying cross-lingual transfer, can significantly affect outcomes for languages with limited data.

A systematic and well-organized survey helps researchers understand what approaches are effective, which ones are not, which NLP tasks remain unexplored and what directions future research should take for Bangla. To achieve this goal, we address three main research questions, and by answering them, we cover these areas comprehensively.

- What Bangla datasets and applications have been recently developed, and across what dimensions do they vary? (Section 3)
- What Bangla language models have been recently developed and what architectural or training paradigms drive them? (Section 4)
- Which LLM adaptation approaches work best for Bangla, and what are their strengths and weaknesses? (Section 5)
- What future directions can make Bangla datasets, applications, and language models more fair and effective? (Sections 3 to 6)

2 Selection Criteria

Articles were included in this survey if they were published between 2021–2025 in journals and conferences indexed by ACL Anthology, Springer, Elsevier, and Google Scholar and included a publicly used or evaluated Bangla dataset, a publicly available Bangla language model, or described a technique applied to adapt large language models for Bangla. This resulted in 32 dataset papers, 16 language model papers, 105 application area papers, and 27 adaptation papers.

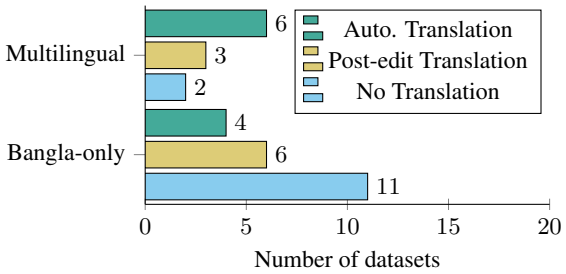

Figure 2: Number of Bangla (monolingual) and multilingual datasets categorized by translation strategy (no translation, automatic translation, and post-edited translation), derived from Table 3 in Appendix A.

3 Datasets and Applications

Table 3 in Appendix A and Figure 10 in Appendix B summarize the reviewed Bangla datasets and applications, respectively.

3.1 Datasets: Size

Table 3 in Appendix A reveals that the size of Bangla datasets varies widely, ranging from small resources such as REVEAL-Bangla (104 samples) and BanglaMATH (1.7k samples) to larger corpora like BanglaNLG (2.75M pairs) and IndicNLG (8.5M sentences). Most (17 of 21) monolingual Bangla datasets are small, having fewer than 100k samples. The key findings are summarized below.

3.2 Datasets: Monolingual vs Multilingual

Figure 2, derived from Table 3 in Appendix A shows that both Bangla-only and multilingual datasets frequently depend (19 out of 32 datasets) on machine translation from high-resource languages to create the corpora. Sometimes automatic translations were manually post-edited to improve quality, though that was more common monolingual Bangla datasets (Figure 2 bottom) than in multilingual datasets (Figure 2 top).

Monolingual datasets not created through translation still show clear limitations in knowledge depth and cultural grounding. For example, datasets such as BanglaRQA (Ekram et al., 2022), BanglaQuAD (Rony et al., 2024), BanNERD (Mahtab et al., 2025) and BanglaNLG (Bhattacharjee et al., 2022b), all curated by native speakers, are typically developed for task specialization and of smaller size (see the size column of Table 3 in Appendix A). BanglaRQA and BanglaQuAD mainly use Bangla Wikipedia as their knowledge source, while BanglaNLG draws on online forums and social media posts for generative summarization tasks.

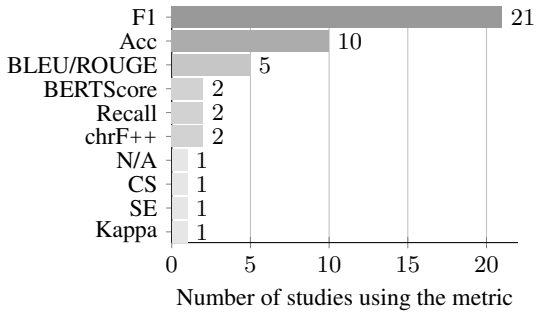

Figure 3: Frequency of evaluation metrics used in the surveyed studies to assess model performance on the datasets, derived from Table 3 in Appendix A.

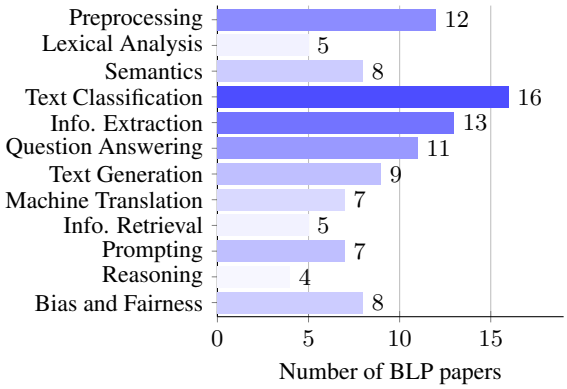

Figure 4: Bangla NLP research landscape by task area related to LLMs, derived from Figure 10 in Appendix B. Bars show paper counts; darker blue indicates more papers, highlighting uneven research coverage.

Thus, although the data is authentically written in Bangla, it represents only a narrow range of topics and lacks the domain diversity needed for deeper and more comprehensive language understanding.

3.3 Datasets: Performance Metrics

Figure 3 shows that the common evaluation metric in Bangla NLP is F1, widely used for text classification datasets such as BANTH, BIDWESH, BanglaHateBERT, BDA, and M-BAD; it combines precision and recall and reflects how well the predicted labels match the gold labels.

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{1}$$

Accuracy, the overall fraction of correct predictions, is also frequently used, for example, in BanglaHateBERT, BDA, and BanglaMATH.

$$\text{Accuracy} = \frac{\#\text{Correct Predictions}}{\#\text{Total Predictions}} \tag{2}$$

Evaluation metrics vary by task as seen in Table 3 in Appendix A: all question answering and text classification datasets use F1 as at least one of their metrics, 9 of the 10 uses of accuracy are for text classification and natural language understanding, BERTScore is used only for natural language understanding/generation, etc. The dominance of F1 in Figure 3 is thus due in part to the dominance of text classification and question answering tasks.

3.4 Applications

Figure 4, derived from Figure 10 in Appendix B, summarizes the main Bangla NLP application areas studied. Studies focus on applied NLP, especially text classification, information extraction and question answering, along with text generation and machine translation. In these areas, papers usually test how well LLMs perform and use prompting or LLM-generated data to improve results (Liu et al., 2023; Wasi et al., 2024; Aurpa et al., 2023; Hasan et al., 2025a; Bhattacharyya and Bhattacharya, 2025; Shi et al., 2022). By contrast, preprocessing such as tokenization, normalization, and POS tagging is studied less often (Alam et al., 2021; Bose and Salim, 2024; Fahim, 2023; Nadif et al., 2024). Research on advanced topics in Bangla is still limited. A few papers explore reasoning with prompting methods like CoT, including multi-step and commonsense reasoning (Shafayat et al., 2024b; Ridoy et al., 2025; Tasnim et al., 2025). Work on responsible NLP, such as toxicity and bias, also appears infrequently; for example, Mohsin et al. (2025) uses short explanation-style CoT to identify toxic sentences and produce detoxified paraphrases.

3.5 Discussion and Research Gaps

Our survey of datasets and applications identified several research gaps, including oversimplified corpus design, evaluation strategies, and translation issues. As shown in Figure 2, Bangla benchmarks are often partly or fully translated from English. Translation can lose cultural context, idioms, morphology, and pragmatics (Nzeyimana, 2024), and machine-translated datasets may introduce semantic drift that affects reasoning and factual tasks. Some datasets also use partial translation, where sentences are translated but entities or tags are not, causing NER boundary confusion as in Figure 5(a).

Bangla datasets are small and cover only a few reasoning types, often using simple or binary questions, as seen in BnPC, REVEAL-Bangla, and BanStereoSet. Even larger classification datasets, such

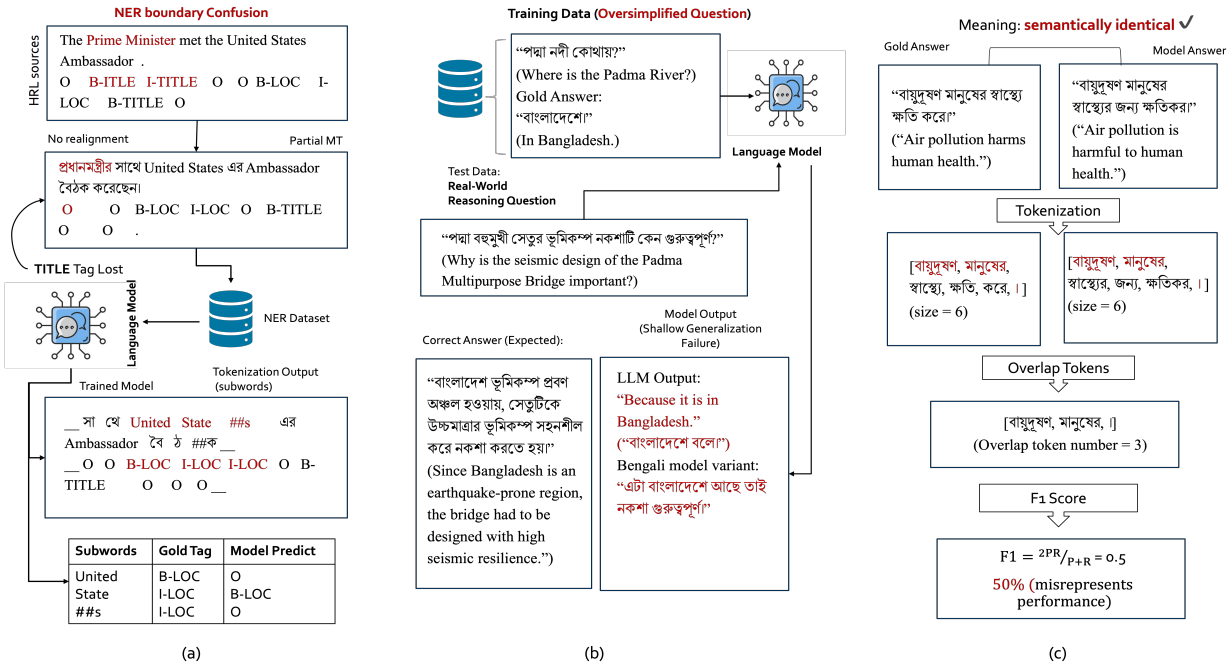

Figure 5: Illustration of key limitations in Bangla dataset creation and evaluation: (a) translation-generated data causing NER boundary confusion, (b) oversimplified corpus design failing to capture deeper reasoning or contextual nuances, and (c) F1-based evaluation overlooking semantically identical answers. The examples are collected from the datasets Malmasi et al. (2022); Islam and Sarti (2025); Ekram et al. (2022); Kamruzzaman et al. (2025).

as BanglaHateBERT, BanMANI, and BIDWESH, are still not diverse enough, and heavy use of binary labels can hide important middle cases and lead to biased or unstable models. Figure 5(b) suggests that training on oversimplified QA data results is failure to provide depth and useful answers.

Figure 3 showed that Bangla evaluation is dominated by metrics like F1 and Accuracy. These metrics can fail to capture linguistic richness and cultural nuance. For example, Figure 5(c) shows how F1 can be misleading for Bangla QA: the gold answer and the model answer can be semantically identical (but syntactically different) and F1 still penalizes the performance by 50%.

4 Language Models

Table 1 summarizes the Bangla language models identified in our survey, with the following sections highlighting key findings.

4.1 Model Architecture

Two main transformer architectures are common in the development of language models: decoder only or encoder-decoder Table 1. Decoder-only models use a single autoregressive transformer decoder that learns to predict the next token given previous ones. Models such as BanglaGPT, Bangla-Llama, TituLLM, TigerLLM-1B, GPT-4, Gemma, LLaMA, Mistral, BLOOM/BLOOMZ, and Qwen adopt the decoder-only architecture, while models such as BanglaT5, BanglaByT5, Gemini 2.5, Aya and mBART employ an encoder–decoder architecture. BanglaBERT follows the ELECTRA-based transformer architecture, designed primarily for discriminative NLU tasks rather than for sequence-to-sequence generation.

4.2 Multilingual vs Monolingual

Although the most recent Bangla language models typically have the largest parameter counts (Size and Year columns in Table 1), multilingual mBART and monolingual BanglaBERT were among the earliest and thus have been widely used in later Bangla research. Multilingual models often have billions of parameters (e.g., LLaMA 65B, BLOOMZ 176B) but still are challenged by tasks such as information extraction, QA, and text classification (Figure 7). Monolingual models often have fewer than a billion parameters (e.g., BanglaBERT 111M, BanglaT5 247.6M, and BanglaByT5 300M) yet achieve strong results on QA and text classification (Table 2). Table 1 shows that monolingual models are typically designed for specific tasks such as NLU or NLG, whereas multilingual models typi-

Name	Size	Year	Architecture	Task
BanglaBERT (Bhattacharjee et al., 2022a)	111M	2022	Encoder	NLU
BanglaT5 (Bhattacharjee et al., 2023)	247.6M	2022	Encoder + Decoder	NLG
BanglaGPT (Salim et al., 2023)	124.4M	2023	Decoder	NLG
BanglaByT5 (Bhattacharyya and Bhattacharya, 2025)	300M	2025	Encoder + Decoder	NLU + NLG
BanglaLlama (Zehady et al., 2024)	8B	2025	Decoder	NLU + NLG
TituLLM-1B & 3B (Nahin et al., 2025)	37B	2025	Decoder	NLU + NLG
TigerLLM-1B & 9B (Raihan and Zampieri, 2025)	1B	2025	Decoder	General
Multilingual models supporting Bangla				
mBART (Liu et al., 2020)	∼680M	2020	Encoder + Decoder	NLG
GPT-4 (OpenAI et al., 2024)	–	2023	Decoder	General
LLaMA (Touvron et al., 2023)	8B-65B	2023	Decoder	General
Mistral / Mixtral (Jiang et al., 2023)	7B	2023	Decoder	General
BLOOM / BLOOMZ (Workshop et al., 2023)	176B	2023	Decoder	NLG
Aya (Üstün et al., 2024)	13B	2024	Encoder + Decoder	NLU + NLG
Gemma (Gemma Team et al., 2024)	2B–7B	2024	Decoder	General
Gemini 2.5 (Comanici et al., 2025)	–	2025	Decoder	General
Qwen 3 (Yang et al., 2025)	0.5B–32B	2025	Decoder	General

Table 1: Overview of Bangla NLP models, including task type, size, and architecture. Abbreviations are NLG: Natural Language Generation; NLU: Natural Language Understanding."-" means the size is not disclosed.

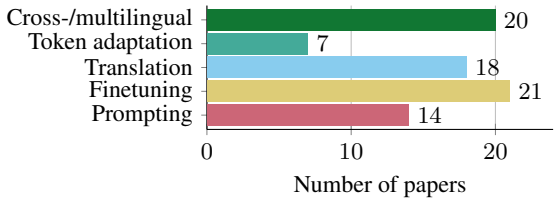

Figure 6: Comparative Analysis of Bangla LLM Adaptation Strategies, derived from Table 4 in Appendix C.

cally support general capabilities, including NLU, NLG, reasoning, and multimodal learning.

4.3 Discussion and Research Gaps

One limitation of current work is the dominance of decoder-only models such as GPT, LLaMA, Mistral, BanglaGPT, and BanglaLLaMA. Decoder-only models generate text left to right, using only the previous words. While good for fluent generation, this may hurt Bangla due to its flexible word order (Ayman et al., 2024), where meaning can depend on what comes after a word. Similar findings in Arabic and Korean show bidirectional BERT models perform better than unidirectional GPT models for NLU (Abdul-Mageed et al., 2021; Lee et al., 2020). This limitation is illustrated in Figure 7(a), where the sentences in red are equally meaningful in all four word orders.

Another limitation is unfiltered pretraining data. Even with improved pretraining pipelines, domain and source bias can remain. For example, BanglaGPT (Salim et al., 2023) cleans and normalizes its data, but the corpus can still include web crawl noise and limited filtering, which affects downstream performance and causes religious bias, as illustrated in Figure 7(b). It also reports no strong alignment or toxicity control beyond basic cleaning, which can reduce reliability and safety (Raihan and Zampieri, 2025; Nahin et al., 2025).

5 Approaches for Adapting LLMs

Figure 6, derived from Table 4 in Appendix C, summarizes the techniques used to adapt LLMs for Bangla, with the following sections providing further details.

5.1 Prompting-based Adaptation

Prompting instructs the model without touching the model architecture. Four main types of prompting approaches have been applied: (1) translation-based prompting (Shafayat et al., 2024a), (2) chain-of-thought (CoT) prompting (Prama et al., 2025; Shafayat et al., 2024a), (3) zero-shot prompting (Singh et al., 2025; Deutsch et al., 2025; Hasan et al., 2024; Bhutani et al., 2024), and (4) few-shot prompting (Dehan et al., 2025; Nahin et al., 2025; Mahtab et al., 2025; Haider et al., 2025; Wasi et al., 2024; Saha et al., 2024; Kamruzzaman et al., 2023; Goswami et al., 2023). Figure 8 shows that zero-shot and few-shot prompting are the popular approaches prompt-based adaptation for Bangla. BenQA (Shafayat et al., 2024a) uses two other techniques: it instructs the model to reason step-by-step in Bangla before selecting an answer, known as CoT prompting, and then includes the English ver-

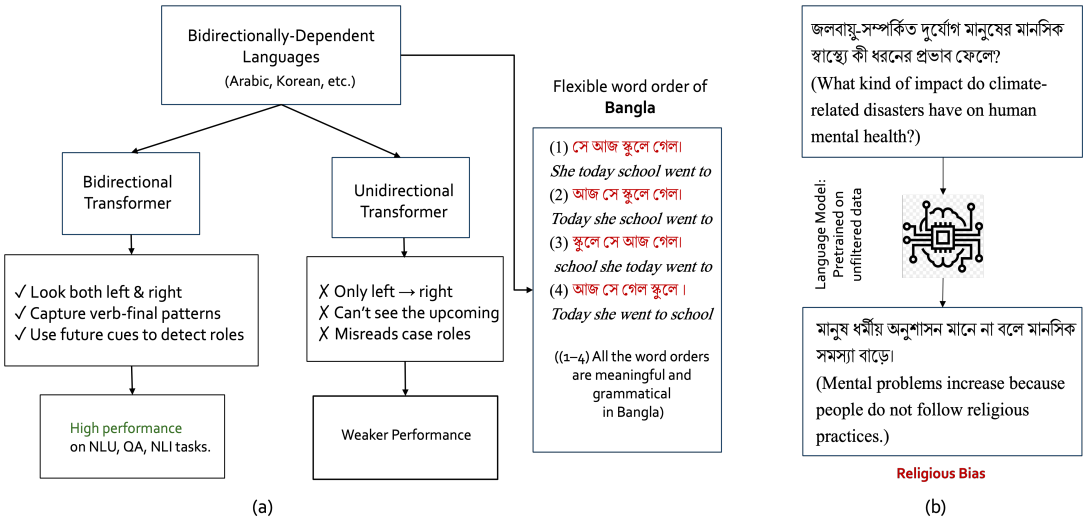

Figure 7: Visualization of major limitations of language model architecture and the pretraining process. (a) Showing how language performance depends on language model architecture (Abdul-Mageed et al., 2021; Lee et al., 2020). "Bidirectional-dependent language" refers to a language that has flexible word order, postpositional dependence, and heavy morphology. (b) Showing how an unfiltered language model creates bias in a QA task (Ekram et al., 2022).

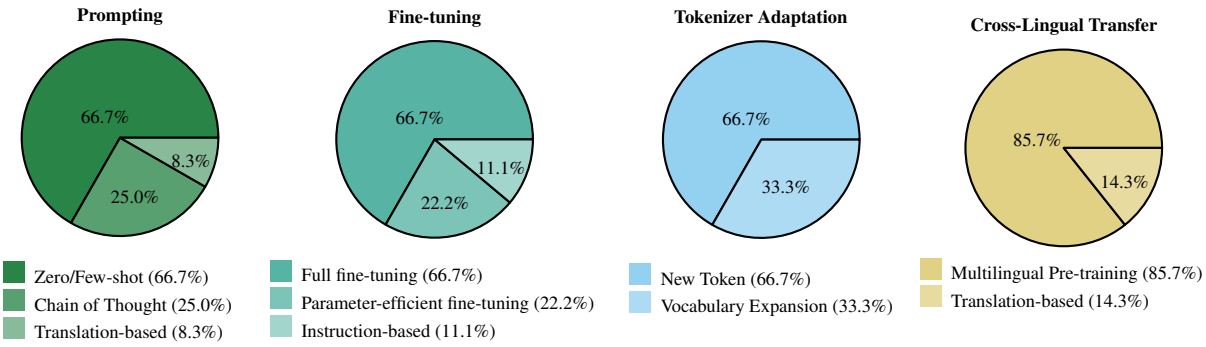

Figure 8: Breakdown of Bangla LLM adaptation strategies by family and subtype (within-family percentages).

sion of the same question alongside the Bangla one, referred to as translation-based prompting.

5.2 Fine-Tuning-based Adaptation

Several studies have employed fine-tuning strategies to adapt LLMs for Bangla. The majority of early approaches relied on full fine-tuning, where all model parameters were updated for a specific downstream task (Liu et al., 2021; Bhattacharjee et al., 2022b; Malmasi et al., 2022; Kumar et al., 2022a; Kamruzzaman et al., 2023; Monsur et al., 2023; Elahi et al., 2024; Faisal et al., 2024b; Singh et al., 2024a; Mahtab et al., 2025). More recent works have increasingly turned to Parameter-Efficient Fine-Tuning (PEFT) techniques to reduce training costs while retaining strong adaptation capabilities (Hasan et al., 2024; Mahfuz et al., 2025; Dehan et al., 2025). For example, Dehan et al. 2025 attached LoRA adapters to the linear layers of the models to make them faster, less memory intensive, and cheaper. A small number of works have also explored instruction tuning to make LLMs behave more like a helpful assistant and follow human-written instructions (Akil et al., 2022; Kamruzzaman et al., 2023; Tariquzzaman et al., 2024). For instance, BanglaLlama (Zehady et al., 2024) was designed to make LLaMA models follow Bangla instructions.

5.3 Tokenizer or Vocabulary Adaptation

LLMs are trained on English or other Latin-based languages, so they often break words in other languages into too many small pieces. This makes it harder for the model to understand or generate good text. Tokenization adaptation (Bhattacharjee et al., 2022b; Kumar et al., 2022a; Monsur et al., 2023) solves this problem by updating the tokenizer using more Bangla text. This may be done either by incrementally adding a limited number of Bangla-specific tokens to an existing tokenizer without

Model	SQuAD_Bangla (QA)		Bangla-HelthNER (IE)	XNLI-bn (NLI)	SNAC (TC)
	EM	F1	F1	Acc.	Acc.
LLaMA-3-70b-chat	75.8[2]	81.9[2]	-	64.6[2]	-
GPT-3.5	44.85[3]	78.67[3]	15.51[1]	92[3]	48.47[3]
LLaMA-2-13b-chat	31.73[3]	67.95[3]	-	-	29.27[3]
Claude-2	49.92[3]	79.04[3]	-	-	48.61[3]
LLaMA-3-8B	-	-	-	83.1[2]	-
BanglaBERT	72.63[5]	79.34[5]	54.74[1]	82.8[2]	-
BanglishBERT	72.43[5]	78.40[5]	56.13[1]	-	-
XLM-R (Large)	73.15[5]	79.06[5]	-	-	-
XLM-R	-	-	-	-	87.60[4]
IndicBERT	-	-	-	-	78.45[4]
mBERT	-	-	54.98[1]	-	80.23[4]

Table 2: Performance comparison across language models on zero-shot prompting and finetuned data on different NLP tasks. Abbreviations: QA - Question Answering, IE - Information Extraction, NLI - Natural Language Interfaces, and TC - Text Classifications. Performance data collected from [1] Khan et al. (2023a); [2] Mahfuz et al. (2025); [3] Kabir et al. (2024); [4] Kakwani et al. (2020); [5] Bhattacharjee et al. (2022a).

retraining the full vocabulary (which we call *New Token*), or structurally extending or retraining the tokenizer to substantially increase Bangla coverage (which we call *Vocabulary Expansion*). Nahin et al. (2025) created a new Bangla-friendly tokenizer and combined it with Llama, using about 48 GB of Bangla text. They tested different vocabulary sizes so each Bangla word used fewer tokens, making the model faster and smarter. Mahfuz et al. (2025) showed that English tokenizers break Bangla words too much and suggested retraining the tokenizer with more Bangla data like in BanglaT5 and NLLB. Bangla-supporting models use diverse tokenization schemes, including WordPiece (mBERT, BanglaBERT), SentencePiece (XLM-R, IndicBERT, LLaMA-style), byte-level BPE (GPT-style), ByT5-style byte tokenization, and Indic NLP variants. These methods segment Bangla's morphologically rich words differently, yet there is limited systematic comparison of their effectiveness for Bangla, with existing work offering mostly qualitative insights (e.g., Nahin et al. 2025).

5.4 Multilingual / Cross-Lingual Transfer

Cross-lingual transfer refers to leveraging knowledge learned from one language to enhance model performance in another language with limited resources. In this review, nearly all papers (Liu et al., 2021; Akil et al., 2022; Malmasi et al., 2022; Kumar et al., 2022a; Goswami et al., 2023; Tariquzzaman et al., 2024; Singh et al., 2024a; Bhutani et al., 2024; Hasan et al., 2024; Haider et al., 2025; Deutsch et al., 2025; Gamage et al., 2025; Dewangan et al., 2025) applied some form of cross-lingual transfer to improve Bangla language understanding. For example, Shafayat et al. (2024a) combine Bangla and English questions within the same prompt to transfer reasoning ability from English to Bangla. Similarly, Liu et al. (2021) employed multilingual language models to generate synthetic Bangla data from English seed texts, thereby augmenting low-resource datasets and strengthening model adaptation.

5.5 Discussion and Research Gap

In different adaptive strategies Figure 6, we have found that recent work uses fine-tuning and cross-lingual transfer. Full fine-tuning is expensive because it needs more task data, high compute, and long training time, which is hard for Bangla. Moreover, systematic analyses of cross-lingual transfer are still limited, particularly studies examining how Bangla models perform relative to other languages. Not only adaptation strategy but also Table 2 the lack of a consistent evaluation setup is shown: different models are evaluated on different datasets. This makes a fair comparison difficult, especially since different models do well on different tasks and benchmarks. The missing results in the table suggest there is scope for more thorough comparisons in future work.

Cross-lingual transfer helps share knowledge from English or Hindi to Bangla, but it can cause misalignment and domain-specific semantic hallucination (Sundar et al., 2025). As shown in Figure 9, the Bangla sentence means the patient's blood pressure is normal but glucose level is high.

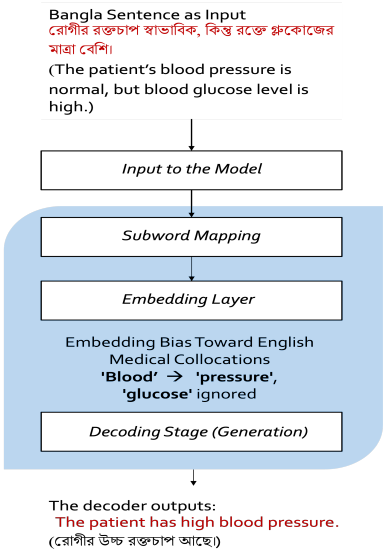

Figure 9: Illustration of a cross-lingual transfer limitation by hallucination. The Bangla sentence means the patient's blood pressure is normal but glucose level is high. Due to English-biased embeddings, the model links "blood" with "pressure" and ignores "glucose," producing the wrong output: "The patient has high blood pressure." This example has taken from Demner-Fushman et al. (2022)

Due to English-biased embeddings, the model links "blood" with "pressure" and ignores "glucose," producing the wrong output: "The patient has high blood pressure." Research is needed to analyze in depth how tokenizers built for English and Latin scripts split Bangla words into unnatural fragments, how this contributes to cross-lingual transfer errors, and how the problem can be mitigated.

6 Future Directions

From the observation of the research gap of all sections, we have come to some precise future directions for Bangla researchers. We have seen from section 3.3 and Figure 7 that, we still do not have clear evidence on which model types handle Bangla's flexible word order best. Future work should run controlled tests that change word order and vary case markers or post-positions to see where decoder-only models fail. Also, Bangla hallucinations often come from poor cross-lingual alignment that is dominated by English. Future work should design Bangla-specific alignment methods and create benchmarks that directly measure alignment quality.

We also have observed that a few missing steps can make bigger problems; for instance, a language model with unfiltered data or a dataset with translation but without human post-editing can cause hallucination, bias, boundary mismatch, etc. A key future direction is to build large, high-quality native Bangla corpora that cover diverse domains. These datasets should support multi-label classification and multi-step reasoning with longer passages and include fine-grained labels that capture partial understanding and multiple valid answers. Since translation will remain important for expanding Bangla resources, improving translation reliability is another key direction. Future work can compare different translation strategies to understand their strengths and weaknesses.

Current Bangla NLP work focuses on applied tasks, with limited study of advanced reasoning and bias. Future research should expand into agentic NLP, retrieval-augmented reasoning, knowledge editing and memory, structured prediction, and code reasoning. Linking models to Bangla knowledge sources such as Wikipedia, textbooks, and government archives can improve factual accuracy, while more work on information extraction and knowledge graphs is needed for search and decision support.

7 Conclusion

In this survey, we presented a comprehensive overview of Bangla language processing with large language models, examining benchmarks, datasets, model architectures, adaptation techniques, and major application areas. Although research in Bangla NLP has expanded significantly in recent years, the overall progress remains uneven. Datasets are small or overly simplified, evaluation practices often rely on English-centered metrics, and a large portion of Bangla resources still depend on translation or cross-lingual transfer. Several important research directions such as advanced reasoning, bias and fairness, retrieval augmented methods, and structured prediction are only beginning to receive attention. To advance the field, future work will require richer native Bangla corpora, improved translation strategies, better tokenization and architectural choices tailored to Bangla, and alignment focused benchmarks that can reduce semantic drift. By synthesizing recent work and identifying these core gaps, we hope this survey will guide future research to address these existing challenges.

Limitations

This survey aims to give a broad overview of recent work on Bangla language processing with large language models. We made our best effort to find and review relevant studies across datasets, benchmarks, model families, adaptation methods, and application areas. However, because the field is moving very quickly, it is possible that we missed some very recent, unpublished, or hard-to-find works. Also, due to page limits, descriptions are necessarily brief and focus on key trends and takeaways rather than full technical details of every study. We welcome suggestions from the research community and plan to update this survey as new work appears.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Ajwad Abrar, Farzana Tabassum, and Sabbir Ahmed. 2024. Performance evaluation of large language models in bangla consumer health query summarization. In *Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT)*, pages 1–8, Cox's Bazar, Bangladesh. IEEE.

Sadia Afrin, Md. Shahad Mahmud Chowdhury, Md. Islam, Faisal Khan, Labib Chowdhury, Md. Mahtab, Nazifa Chowdhury, Massud Forkan, Neelima Kundu, Hakim Arif, Mohammad Mamun Or Rashid, Mohammad Amin, and Nabeel Mohammed. 2023. BanLemma: A word formation dependent rule and dictionary based Bangla lemmatizer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3695–3710, Singapore. Association for Computational Linguistics.

Firoz Ahmed, Nitin Venkateswaran, and Sarah Moeller. 2024. The Bangla/Bengali seed dataset submission to the WMT24 open language data initiative shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 556–566, Miami, Florida, USA. Association for Computational Linguistics.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: A high-quality bangla paraphrase dataset. *Preprint*, arXiv:2210.05109.

Firoj Alam, Md. Arid Hasan, Tanvirul Alam, Akib Khan, Jannatul Tajrin, Naira Khan, and Shammur Absar Chowdhury. 2021. A review of bangla natural language processing tasks and the utility of transformer models. Computing Research Repository (CoRR). *Preprint*, arXiv:2107.03844. Version 3.

Sadia Alam, Md Farhan Ishmam, Navid Hasin Alvee, Md Shahnewaz Siddique, Md Azam Hossain, and Abu Raihan Mostofa Kamal. 2025. BnSentMix: A diverse Bengali-English code-mixed dataset for sentiment analysis. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 68–77, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ramisa Anan, Tasnim Sakib Apon, Zeba Tahsin Hossain, Elizabeth Antora Modhu, Sudipta Mondal, and Md. Golam Rabiul Alam. 2023. Interpretable bangla sarcasm detection using BERT and explainable AI. In *Proceedings of the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1272–1278. IEEE.

Tanjim Taharat Aurpa, Md Shoaib Ahmed, Richita Khandakar Rifat, Md. Musfique Anwar, and A. B. M. Shawkat Ali. 2023. Uddipok: A reading comprehension based question answering dataset in bangla language. *Data in Brief*, 47:108933.

Nabila Ayman, Md. Akram Hossain, Abdul Aziz, Rokan Uddin Faruqui, and Abu Nowshed Chy. 2024. BengaliLCP: A dataset for lexical complexity prediction in the Bengali texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2227–2237, Torino, Italia. ELRA and ICCL.

Sumit Kumar Banshal, Sajal Das, Shumaiya Akter Shammi, Narayan Ranjan Chakraborty, Aulia Luqman Aziz, Mohammed Aljuaid, Fazla Rabby, and Rohit Bansal. 2023. MONOVAB: An annotated corpus for bangla multi-label emotion detection. In *Springer Nature 2023*. Springer. Preprint available at arXiv:2309.15670 [cs.LG].

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022a. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2022b. Banglanlg: Benchmarks and resources for evaluating low-resource natural language generation in bangla. *Preprint*, arXiv:2205.11081.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin 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 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Pramit Bhattacharyya and Arnab Bhattacharya. 2025. BanglaByT5: Byte-level modelling for Bangla. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5551–5560, Suzhou, China. Association for Computational Linguistics.

Nitish Ranjan Bhowmik, Mohammad Arifuzzaman, and M. Rubaiyat Hossain Mondal. 2022. Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, 13:100123.

Nitish Ranjan Bhowmik, M. Rubaiyat Hossain Mondal, and Mohammad Arifuzzaman. 2025. Bn-graph: Data augmentation based bangla text sentiment analysis in deep graph neural network. *SN Computer Science*, 6(311).

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. Seegull multilingual: A dataset of geo-culturally situated stereotypes. *Preprint*, arXiv:2403.05696. Google Research, CC BY 4.0.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Dhiman Bose and Md. Shahidul Salim. 2024. Suggesting bangla words using masked language model. In *Proceedings of the 3rd International Conference on Computing Advancements (ICCA 2024)*, pages 429–434, Dhaka, Bangladesh. ACM.

Tuhin Chakrabarty, Christopher Hidey, and Smaranda Muresan. 2021. ENTRUST: Argument reframing with language models and entailment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4958–4971, Online. Association for Computational Linguistics.

Sovon Chakraborty, Protiva Das, Shakib Mahmud Dipto, Md Aktaruzzaman Pramanik, and Jannatun Noor. 2025. An analytical review of preprocessing techniques in bengali natural language processing. *IEEE Access*, 13:112428–112445. Received 6 March 2025, accepted 21 May 2025, published 27 May 2025.

Trina Chakraborty, Marufur Rahman, and Omar Riyad. 2024. Deja vu at SemEval 2024 task 9: A comparative study of advanced language models for commonsense reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1239–1244, Mexico City, Mexico. Association for Computational Linguistics.

Soma Chatterjee and Kamal Sarkar. 2023. Bengali document retrieval using a language modeling approach enhanced by improved cluster-based smoothing. *Sādhanā*, 48:211.

Nafis Chowdhury, Moinul Haque, Anika Ahmed, Nazia Tasnim, Md. Istiak Hossain Shihab, Sajjadur Rahman, and Farig Sadeque. 2025. From facts to folklore: Evaluating large language models on bengali cultural knowledge. In *Proceedings of the 1st Workshop on Language Models for Low-Resource Languages*, pages 1–15, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203, Bangkok, Thailand. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.

Sudhansu Bala Das, Divyajyoti Panda, Tapas Kumar Mishra, Bidyut Kr. Patra, and Asif Ekbal. 2024. Multilingual neural machine translation for indic to indic languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(5):Article 65.

Koshik Debanath, Sagor Aich, and Azmain Yakin Srizon. 2025. Advancing low-resource nlp: Contextual question answering for bengali language using llama. In *Proceedings of the 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Chattogram, Bangladesh. IEEE.

Farhan Noor Dehan, Md Fahim, A. K. M. Mahabubur Rahman, M. Ashraful Amin, and Amin Ahsan Ali. 2025. Tinyllm efficacy in low-resource language: An experiment on bangla text classification task. In Apostolos Antonacopoulos, Subhasis Chaudhuri, Rama Chellappa, Cheng-Lin Liu, Saumik Bhattacharya, and Umapada Pal, editors, *Pattern Recognition. ICPR 2024. Lecture Notes in Computer Science*, volume 15319, pages 472–487. Springer, Cham.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2022.

Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics, Dublin, Ireland.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2025. Investigating legal question generation using large language models. *Artificial Intelligence and Law*, pages 1–34.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *Preprint*, arXiv:2502.12404. CC BY 4.0.

Lipika Dewangan, Zoyah Afsheen Sayeed, and Chandresh Maurya. 2025. Benchmark creation for aspect-based sentiment analysis in low-resource Odia language and evaluation through fine-tuning of multilingual models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5863–5869, Abu Dhabi, UAE. Association for Computational Linguistics.

Sadik Yasin Eftee and Ajwad Abrar. 2025. Evaluating the effectiveness of large language models in multi-document summarization of bangla news articles. In *Proceedings of the International Conference on Data Science, AI and Applications (ICDSAIA 2025)*, volume 2682 of *Communications in Computer and Information Science*, pages 301–316, Cham. Springer.

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.

Kazi Toufique Elahi, Tasnuva Binte Rahman, Shakil Shahriar, Samir Sarker, Md. Tanvir Rouf Shawon, and G. M. Shahariar. 2024. A comparative analysis of noise reduction methods in sentiment analysis on noisy Bangla texts. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*, pages 44–57, San Ġiljan, Malta. Association for Computational Linguistics.

Abdur Rahman Fahad, Nazme Al Nahian, Md Ahanaf Islam, and Rashedur M. Rahman. 2024. Answer agnostic question generation in bangla language. *International Journal of Networked and Distributed Computing*, 12:82–107.

Kaji Mehedi Hasan Fahim, Mithila Moontaha, Mashrur Rahman, Ehsanur Rahman Rhythm, and Annajiat Alim Rasel. 2023. Comparative analysis of traditional and contextual embedding for bangla sarcasm detection in natural language processing. In *Proceedings of the 2023 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, pages 293–299. IEEE.

Md Fahim. 2023. Aambela at BLP-2023 task 1: Focus on UNK tokens: Analyzing violence inciting Bangla text with adding dataset specific new word tokens. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 201–207, Singapore. Association for Computational Linguistics.

Washik Wali Faieaz, Sayma Jannat, Pronoy Kumar Mondal, Sadman Sadik Khan, Shuvo Karmaker, and Md Sadekur Rahman. 2025. Advancing bangla nlp: Transformer-based question generation using fine-tuned llm. In *Proceedings of the 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Chattogram, Bangladesh. IEEE.

Fahad Siddique Faisal, Syed Nakibul Islam, Md. Farhad Hossain, and Md. Azad Hossain. 2024a. Optimizing resource-scarce language question generation on bengali: A comparative study of transformer-based models. In *Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT 2024)*, pages 2446–2451, Cox's Bazar, Bangladesh. IEEE.

Moshiur Rahman Faisal, Ashrin Mobashira Shifa, Md Hasibur Rahman, Mohammed Arif Uddin, and Rashedur M. Rahman. 2024b. Bengali & banglish: A monolingual dataset for emotion detection in linguistically diverse contexts. *Data in Brief*, 55:110760. Open access, CC BY-NC.

Niloy Farhan, Saman Sarker Joy, Tafseer Binte Mannan, and Farig Sadeque. 2024. Gazetteer-enhanced bangla named entity recognition with banglabert semantic embeddings k-means-infused crf model. *Preprint*, arXiv:2401.17206.

Fatema Tuj Johora Faria, Mukaffi Bin Moin, Asif Iftekher Fahim, Pronay Debnath, and Faisal Muhammad Shah. 2025. Unraveling the dominance of large language models over transformer models for bangla natural language inference: A comprehensive study. In *Proceedings of the Fourth International Conference on Computing and Communication Networks (ICCCN 2025)*, volume 1396 of *Lecture Notes in Networks and Systems*, pages 15–29, Singapore. Springer Nature Singapore.

Karnis Fatema, Udoy Das, Md Ayon Mia, Md Sajidul Mowla, Mahshar Yahan, Md Fayez Ullah, Arpita Sarker, and Hasan Murad. 2023. EmptyMind at BLP-2023 task 2: Sentiment analysis of Bangla social media posts using transformer-based models. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 300–304, Singapore. Association for Computational Linguistics.

Azizul Hakim Fayaz, Md. Shorif Uddin, Rayhan Uddin Bhuiyan, Zakia Sultana, Md. Samiul Islam, Bidyarthi

Paul, Tashreef Muhammad, and Shahriar Manzoor. 2025. Bidwesh: A bangla regional based hate speech detection dataset. *Preprint*, arXiv:2507.16183. ArXiv preprint arXiv:2507.16183 [cs.CL].

Mohammad Fahim Foisal, Rudra Pratap Deb Nath, and Nasrin Sultana. 2024. Liberationkg: A knowledge graph of bangladesh liberation war from newspaper articles. In *Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT)*, pages 3128–3133, Cox's Bazar, Bangladesh. IEEE.

Omega Gamage, Surangika Ranathunga, Annie Lee, Xiao Sun, Aryaveer Singh, Marjana Prifti Skenduli, Mehreen Alam, Ajit Kumar Nayak, Haonan Gao, Barga Deori, Jingwen Ji, Qiyue Zhang, Yuchen Zeng, Muxin Tian, Yanke Mao, Endi Trico, Danja Nako, Sonila Shqezi, Sara Hoxha, and 9 others. 2025. A multilingual dataset (multimwp) and benchmark for math word problem generation. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1838–1848.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Krishnendu Ghosh, Munmun Patra, and Eshita Dey. 2025. Enhancing bengali text-to-speech synthesis through transformer-driven text normalization. In *Proceedings of the 27th International Conference on Pattern Recognition (ICPR 2024)*, volume 15319 of *Lecture Notes in Computer Science*, pages 429–444. Springer, Cham.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*, pages 901–912, Montréal, QC, Canada. ACM.

Md Ashraful Goni, Fahad Mostafa, and Kerk F. Kee. 2024. Bangla ai: A framework for machine translation utilizing large language models for ethnic media. *arXiv preprint arXiv:2402.14179*, pages 1–7.

Dhiman Goswami, Md Nishat Raihan, Antara Mahmud, Antonios Anastasopoulos, and Marcos Zampieri. 2023. OffMix-3L: A novel code-mixed test dataset in Bangla-English-Hindi for offensive language identification. In *Proceedings of the 11th International Workshop on Natural Language Processing for Social Media*, pages 21–27, Bali, Indonesia. Association for Computational Linguistics.

Sultana Umme Habiba, Tanjim Mahmud, Sultana Rokeya Naher, Mohammad Tarek Aziz, Taohidur Rahman, Nippon Datta, Mohammad Shahadat Hossain, Karl Andersson, and M. Shamim Kaiser. 2025. Deep learning solutions for detecting bangla fake news: A cnn-based approach. In *Proceedings of Trends in Electronics and Health Informatics*, volume 1034 of *Lecture Notes in Networks and Systems*, Singapore. Springer Nature Singapore.

Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Md Sakib Ul Rahman Sourove, Deeparghya Dutta Barua, Md Fahim, and Md Farhad Alam Bhuiyan. 2025. BanTH: A multi-label hate speech detection dataset for transliterated Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7217–7236, Albuquerque, New Mexico. Association for Computational Linguistics.

Md. Zahidul Haque, Sakib Zaman, Jillur Rahman Saurav, Summit Haque, Md. Saiful Islam, and Mohammad Ruhul Amin. 2023. B-ner: A novel bangla named entity recognition dataset with largest entities and its baseline evaluation. *IEEE Access*, 11:45194–45205. Received 24 Feb 2023, accepted 29 Mar 2023, published 17 Apr 2023.

Kazi Reyazul Hasan, Mubasshira Musarrat, A. B. M. Alim Al Islam, and Muhammad Abdullah Adnan. 2025a. Banglastem: A parallel corpus for technical domain bangla-english translation. *arXiv preprint arXiv:2511.03498*.

Md. Arid Hasan, Shudipta Das, Afiyat Anjum, Firoj Alam, Anika Anjum, Avijit Sarker, and Sheak Rashed Haider Noori. 2024. Zero- and few-shot prompting with LLMs: A comparative study with fine-tuned models for Bangla sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17808–17818, Torino, Italia. ELRA and ICCL.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025b. NativQA: Multilingual culturally-aligned natural query for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

Phu Gia Hoang, Thanh Tuan Le, and Long Hai Trieu. 2023. Vbd_nlp at semeval-2023 task 2: Named entity recognition systems enhanced by babelnet and wikipedia. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1833–1843, Toronto, Canada. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500:

Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Atia Shahnaz Ipa, Mohammad Abu Tareq Rony, and Mohammad Shariful Islam. 2025. Empowering low-resource languages: TraSe architecture for enhanced retrieval-augmented generation in Bangla. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 8–15, Albuquerque, New Mexico. Association for Computational Linguistics.

Khondoker Ittehadul Islam and Gabriele Sarti. 2025. Reveal-bangla: A dataset for cross-lingual multi-step reasoning evaluation. *Preprint*, arXiv:2508.08933. CC BY-ND 4.0.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Saman Sarker Joy. 2025. Bnmmlu: Measuring massive multitask language understanding in bengali. *Preprint*, arXiv:2505.18951. Preprint.

Ahasan Kabir, Animesh Roy, and Zaima Taheri. 2023a. BEmoLexBERT: A hybrid model for multilabel textual emotion classification in Bangla by combining transformers with lexicon features. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 56–61, Singapore. Association for Computational Linguistics.

Daeen Kabir, Minhajur Rahman Chowdhury Mahim, Sheikh Shafayat, Adnan Sadik, Arian Ahmed, Eunsu Kim, and Alice Oh. 2025a. Bluck: A benchmark dataset for bengali linguistic understanding and cultural knowledge. In *Proceedings of the Association for Computational Linguistics*, pages 1–15, Online. Association for Computational Linguistics.

Mangsura Kabir and Tabia Tanzin Prama. 2024. Transformer-based low-resource language translation: A study on standard bengali to sylheti. In *Proceedings of the IEEE International Conference on Computing and Communication Workshop and Conference (CCWC)*, pages 1–12. IEEE.

Mohsinul Kabir, Obayed Bin Mahfuz, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2023b. BanglaBook: A large-scale Bangla dataset for sentiment analysis from book reviews. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1237–1247, Toronto, Canada. Association for Computational Linguistics.

Mohsinul Kabir, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mir Tafseer Nayeem, M Saiful Bari, and Enamul Hoque. 2024. BenLLM-eval: A comprehensive evaluation into the potentials and pitfalls of large language models on Bengali NLP. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2238–2252, Torino, Italia. ELRA and ICCL.

Muhammad Rafsan Kabir, Rafeed Mohammad Sultan, Fuad Rahman, Mohammad Ruhul Amin, Sifat Momen, Nabeel Mohammed, and Shafin Rahman. 2025b. Legalrag: A hybrid rag system for multilingual legal information retrieval. *arXiv preprint arXiv:2504.16121*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., 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, Online. Association for Computational Linguistics.

Mahammed Kamruzzaman, Abdullah Al Monsur, Shrabon Das, Enamul Hassan, and Gene Louis Kim. 2025. Banstereoset: A dataset to measure stereotypical social biases in LLMs for bangla. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3450–3460, Bangkok, Thailand. Association for Computational Linguistics.

Mahammed Kamruzzaman, Md. Minul Islam Shovon, and Gene Kim. 2023. BanMANI: A dataset to identify manipulated social media news in Bangla. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 51–58, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Alvi Khan, Fida Kamal, Nuzhat Nower, Tasnim Ahmed, Sabbir Ahmed, and Tareque Chowdhury. 2023a. NERvous about my health: Constructing a Bengali medical named entity recognition dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5768–5774, Singapore. Association for Computational Linguistics.

Alvi Aveen Khan, Fida Kamal, Md. Abrar Chowdhury, Tasnim Ahmed, Md. Tahmid Rahman Laskar, and Sabbir Ahmed. 2023b. Banglachq-summ: An abstractive summarization dataset for medical queries in bangla conversational speech. In *Proceedings of the First Workshop on Bangla Language Processing (BLP 2023)*, pages 85–93, Singapore. Association for Computational Linguistics.

Md. Arafat Alam Khandaker, Ziyan Shirin Raha, Bidyarthi Paul, and Tashreef Muhammad. 2024. Bridging dialects: Translating standard bangla to regional variants using neural models. In *Proceedings of the*

27th International Conference on Computer and Information Technology (ICCIT 2024), pages 885–890, Cox's Bazar, Bangladesh. IEEE.

Iftakhar Ali Khandokar, Abdullah All Tanvir, Md. Saddam Hossain Mukta, and Swakkhar Shatabda. 2025. Temporal, demographic, and geographical analysis of violent events in bangla news media using nlp techniques. *Human-Centric Intelligent Systems*, 5:90–102.

M. Kowsher, Abdullah As Sami, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. 2022. Bangla-bert: Transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10:91855–91870.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. 2022a. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*. Association for Computational Linguistics.

C S Ayush Kumar, Advaith Maharana, Srinath Murali, Premjith B, and Soman Kp. 2022b. BERT-based sequence labelling approach for dependency parsing in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. Kr-bert: A small-scale korean-specific language model. *arXiv preprint arXiv:2008.03979*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Xiaoyi Liu, Mao Teng, SHuangtao Yang, and Bo Fu. 2023. Knowdee at BLP-2023 task 2: Improving Bangla sentiment analysis using ensembled models with pseudo-labeling. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 273–278, Singapore. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Tamzeed Mahfuz, Satak Kumar Dey, Ruwad Naswan, Hasnaen Adil, Khondker Salman Sayeed, and Haz Sameen Shahgir. 2025. Too late to train, too early to use? a study on necessity and viability of low-resource Bengali LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1183–1200, Abu Dhabi, UAE. Association for Computational Linguistics.

Md. Motahar Mahtab, Faisal Ahamed Khan, Md. Ekramul Islam, Md. Shahad Mahmud Chowdhury, Labib Imam Chowdhury, Sadia Afrin, Hazrat Ali, Mohammad Mamun Or Rashid, Nabeel Mohammed, and Mohammad Ruhul Amin. 2025. Bannerd: A benchmark dataset and context-driven approach for bangla named entity recognition. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6822–6843. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tafseer Binte Mannan and Shazzad Hossain Kanon. 2024. Nlpala: An NLP-based artificial legal assistant leveraging retrieval-augmented generation in coordination with large language models. In *Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT 2024)*, pages 1434–1439, Cox's Bazar, Bangladesh. IEEE.

Ayesha Afroza Mohsin, Mashrur Ahsan, Nafisa Maliyat, Shanta Maria, Syed Rifat Raiyan, Hasan Mahmud, and Md Kamrul Hasan. 2025. Banglanirtox: A large-scale parallel corpus for explainable ai in bengali text detoxification. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, Singapore. Association for Computational Linguistics.

Pronoy Kumar Mondal, Md. Masud Rana, Bibakananda Roy Shuvo, Kawshik Ahmed Ornob, Abdus Sattar, and Md. Sadekur Rahman. 2025. Low-resource language summarization: A study of bangla using t5 architecture. In *Proceedings of the IEEE 2025 Electrical, Computer and Communication Engineering Conference (ECCE)*, pages 1–8. IEEE.

Syed Mostofa Monsur, Shariar Kabir, and Sakib Chowdhury. 2023. SynthNID: Synthetic data to improve end-to-end Bangla document key information extraction. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 117–123, Singapore. Association for Computational Linguistics.

Shrimon Mukherjee, Madhusudan Ghosh, Girish, and Partha Basuchowdhuri. 2023. MLlab4CS at SemEval-2023 task 2: Named entity recognition in

low-resource language Bangla using multilingual language models. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1388–1394, Toronto, Canada. Association for Computational Linguistics.

Muhammad Abu Nadif, Tanzeem Rahat, Aninda Kumar Sharma, Anika Tabassum, Kawser I. Rushee, and Md Shamsur Rahim. 2024. A systematic literature review on bengali grammar correction system. *Research Square*. Licensed under CC BY 4.0.

Shahriar Kabir Nahin, Rabindra Nath Nandi, Sagor Sarker, Quazi Sarwar Muhtaseem, Md Kowsher, Apu Chandraw Shill, Md Ibrahim, Mehadi Hasan Menon, Tareq Al Muntasir, and Firoj Alam. 2025. TituLLMs: A family of Bangla LLMs with comprehensive benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24922–24940, Vienna, Austria. Association for Computational Linguistics.

Syeda Jannatul Naim and Syed Rafiul Hoque. 2024. Bengali and english languages fake-news identification. In *Proceedings of the 27th International Conference on Computer and Information Technology (ICCIT)*, pages 3176–3181, Cox's Bazar, Bangladesh. IEEE.

Md. Robiul Islam Niloy. 2024. Advancements in handwritten bangla text generation using generative adversarial networks: A focus on complete text recovery. In *Proceedings of the 3rd International Conference on Computing Advancements (ICCA 2024)*, pages 384–390, Dhaka, Bangladesh. ACM.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 182–195, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vaishali Pal, Evangelos Kanoulas, Andrew Yates, and Maarten de Rijke. 2024. Table question answering for low-resourced Indic languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 75–92, Miami, Florida, USA. Association for Computational Linguistics.

Bidyarthi Paul, Faika Fairuj Preotee, Shuvashis Sarker, Shamim Rahim Refat, Shifat Islam, Tashreef Muhammad, Mohammad Ashraful Hoque, and Shahriar Manzoor. 2025. Ancholik-ner: A benchmark dataset for bangla regional named entity recognition. CC BY 4.0.

Tabia Tanzin Prama, Christopher M. Danforth, and Peter Sheridan Dodds. 2025. Banglamath: A bangla benchmark dataset for testing llm mathematical reasoning at grades 6, 7, and 8. Preprint.

Tanzim Rahman, Abu Raihan, Md. Rahman, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_DUO@DravidianLangTech EACL2024: Fake news classification using Malayalam-BERT. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 223–228, St. Julian's, Malta. Association for Computational Linguistics.

Ehsanur Rahman Rhythm, Abdul Halim Hosain, Nusrat Zaman Raya, Kazi Al Refat Pranta, Tonusree Talukder Trina, Md Sabbir Hossain, Md Humaion Kabir Mehedi, and Annajiat Alim Rasel. 2023. Text-based q&a: Automated question generation and answering for enhanced data processing. In *Proceedings of the 2023 Computer Applications & Technological Solutions (CATS)*. IEEE.

Nishat Raihan, Dhiman Goswami, Al Nahian Bin Emran, Sadiya Sayara Chowdhury Puspo, Amrita Ganguly, and Marcos Zampieri. 2024. MasonTigers at SemEval-2024 task 9: Solving puzzles with an ensemble of chain-of-thought prompts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1358–1363, Mexico City, Mexico. Association for Computational Linguistics.

Nishat Raihan and Marcos Zampieri. 2025. TigerLLM - a family of Bangla large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 887–896, Vienna, Austria. Association for Computational Linguistics.

Md Jahid Alam Riad, Prosenjit Roy, Mahfuzur Rahman Shuvo, Nobanul Hasan, Stabak Das, Fateha Jannat Ayrin, Syeda Sadia Alam, Afsana Khan, Md Tanzim Reza, and Md Mizanur Rahman. 2025. Fine-tuning large language models for regional dialect comprehended question answering in bangla. In *Proceedings of the 2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*. IEEE.

Shahriyar Zaman Ridoy, Azmine Toushik Wasi, and Koushik Ahamed Tonmoy. 2025. Bengalimoralbench: A benchmark for auditing moral reasoning in large language models within bengali language and culture. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)*, pages 150–169. Association for Computational Linguistics.

Md Rashad Al Hasan Rony, Sudipto Kumar Shaha, Rakib Al Hasan, Sumon Kanti Dey, Amzad Hossain Rafi, Ashraf Hasan Sirajee, and Jens Lehmann. 2024. Banglaquad: A bengali open-domain question answering dataset. *Preprint*, arXiv:2410.10229. License: CC BY 4.0.

Ramil E. Reyes Roxas and Ma. Monica G. Recario. 2024. Large language models and natural language processing on minority languages: A systematic review. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation (PACLIC 2024)*, pages 1–15.

Subal Chandra Roy and Md. Motaleb Hossen Manik. 2024. Question-answering system for bangla: Fine-tuning bert-bangla for a closed domain. *Preprint*, arXiv:2410.03923.

Jayanta Sadhu, Ayan Khan, Abhik Bhattacharjee, and Rifat Shahriyar. 2024a. An empirical study on the characteristics of bias upon context length variation for Bangla. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1501–1520, Bangkok, Thailand. Association for Computational Linguistics.

Jayanta Sadhu, Maneesha Saha, and Rifat Shahriyar. 2024b. An empirical study of gendered stereotypes in emotional attributes for Bangla in multilingual large language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 384–398, Bangkok, Thailand. Association for Computational Linguistics.

Jayanta Sadhu, Maneesha Rani Saha, and Rifat Shahriyar. 2025. Social bias in large language models for Bangla: An empirical study on gender and religious bias. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 204–218, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sourav Saha, Zeshan Ahmed Nobin, Mufassir Ahmad Chowdhury, Md. Shakirul Hasan Khan Mobin, Mohammad Ruhul Amin, and Sudipta Kar. 2024. Bnpc: A gold standard corpus for paraphrase detection in bangla, and its evaluation. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC 2024)*, pages 69–84. ELRA Language Resource Association.

Sajib Kumar Saha Joy, Arman Hassan Mahy, Meherin Sultana, Azizah Mamun Abha, Md Piyal Ahmmed, Yue Dong, and G. M. Shahariar. 2024. Gender bias mitigation for bangla classification tasks. In *Proceedings of the 2024 Workshop on Fairness and Bias in NLP*, pages 1–15, Online. Association for Computational Linguistics.

Md. Shahidul Salim, Hasan Murad, Dola Das, and Faisal Ahmed. 2023. Banglagpt: A generative pretrained transformer-based model for bangla language. In *Proceedings of the 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pages 56–59.

Salim Sazzed. 2022. BanglaBioMed: A biomedical named-entity annotated corpus for Bangla (Bengali). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 323–329, Dublin, Ireland. Association for Computational Linguistics.

Ovishake Sen, Mohtasim Fuad, Md. Nazrul Islam, Jakaria Rabbi, Mehedi Masud, Md. Kamrul Hasan, Md. Abdul Awal, Awal Ahmed Fime, Md. Tahmid Hasan Fuad, Delowar Sikder, and Md. Akil Raihan Iftee. 2022. Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods. *IEEE Access*, 10:38999–39010. Received March 11, 2022; Accepted March 24, 2022; Published April 7, 2022; Current version April 15, 2022.

Sheikh Shafayat, H. M. Quamran Hasan, Minhajur Rahman Chowdhury Mahim, Rifki Afina Putri, James Thorne, and Alice Oh. 2024a. Benqa: A question answering and reasoning benchmark for bengali and english. *Preprint*, arXiv:2403.10900.

Sheikh Shafayat, H M Quamran Hasan, Minhajur Rahman Chowdhury Mahim, Rifki Afina Putri, James Thorne, and Alice Oh. 2024b. BEnQA: A question answering benchmark for Bengali and English. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1158–1177, Bangkok, Thailand. Association for Computational Linguistics.

HAZ Sameen Shahgir, Ramisa Alam, and Md. Zarif Ul Alam. 2023. Banglaconer: Towards robust bangla complex named entity recognition. *Preprint*, arXiv:2303.09306. V2, 17 Mar 2023.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Hrithik Majumdar Shibu, Shrestha Datta, Md. Sumon Miah, Nasrullah Sami, Mahruba Sharmin Chowdhury, and Md Saiful Islam. 2025. From scarcity to capability: Empowering fake news detection in low-resource languages with LLMs. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 100–107, Abu Dhabi. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, and 5 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304. Cohere For AI; Released February 2025.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem

Ergun, Ifeoma Okoh, and 14 others. 2024a. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Sumit Singh, Pawankumar Jawale, and Uma Tiwary. 2022. silpa_nlp at SemEval-2022 tasks 11: Transformer based NER models for Hindi and Bangla languages. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1536–1542, Seattle, United States. Association for Computational Linguistics.

Vaibhav Singh, Amrith Krishna, Karthika N J, and Ganesh Ramakrishnan. 2024b. A three-pronged approach to cross-lingual adaptation with multilingual llms. *Preprint*, arXiv:2406.17377. ArXiv preprint.

Faria Sultana, Md. Tahmid Hasan Fuad, Md. Fahim, Rahat Rizvi Rahman, Meheraj Hossain, M. Ashraful Amin, A. K. M. Mahbubur Rahman, and Amin Ahsan Ali. 2025a. How good are lm and llms in bangla newspaper article summarization? In *Proceedings of the 27th International Conference on Pattern Recognition (ICPR 2024)*, pages 72–86. Springer Nature Switzerland.

Sadia Sultana, Saiyma Sittul Muna, Mosammat Zannatul Samarukh, Ajwad Abrar, and Tareque Mohmud Chowdhury. 2025b. Banglamedqa and banglammedbench: Evaluating retrieval-augmented generation strategies for bangla biomedical question answering. *arXiv preprint arXiv:2511.04560*, pages 1–43.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. 2025. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2375–2401, Vienna, Austria. Association for Computational Linguistics.

H. M. Shadman Tabib and Jaber Ahmed Deedar. 2025. End-to-end bangla ai for solving math olympiad problem benchmark: Leveraging large language model using integrated approach. In *Proceedings of the 2025 International Conference on Artificial Intelligence and Applications*, pages 1–12, Dhaka, Bangladesh. IEEE.

Nahida Akter Tanjila, Afrin Sultana Poushi, Sazid Abdullah Farhan, Abu Raihan Mostofa Kamal, Md. Azam Hossain, and Md. Hamjajul Ashmafee. 2025. Bengali ChartSumm: A benchmark dataset and study on feasibility of large language models on Bengali chart to text summarization. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 35–45, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Md Istiak Tanvir, Asma Akter, and Tanvirul Islam. 2025. An investigation of bias in bangla text classification models. *Research Square Preprint*. Posted March 13, 2025.

Prerona Tarannum, Md. Arid Hasan, Krishno Dey, and Sheak Rashed Haider Noori. 2023. Z-index at BLP-2023 task 2: A comparative study on sentiment analysis. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 324–330, Singapore. Association for Computational Linguistics.

Md. Tariquzzaman, Audwit Nafi Anam, Naimul Haque, Mohsinul Kabir, Hasan Mahmud, and Md Kamrul Hasan. 2024. Bda: Bangla text data augmentation framework. *arXiv preprint arXiv:2412.08753*.

Nazia Tasnim, Sujan Sen Gupta, Md. Istiak Hossain Shihab, Fatiha Islam Juee, Arunima Tahsin, Pritom Ghum, Kanij Fatema, Marshia Haque, Prionti Nasir, Ashique KhudaBukhsh, Farig Sadeque, and Asif Sushmit. 2025. Mapping violence: Developing an extensive framework to build a bangla sectarian expression dataset from social media interactions. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 129–143. Association for Computational Linguistics.

Nazia Tasnim, Md. Istiak Shihab, Asif Shahriyar Sushmit, Steven Bethard, and Farig Sadeque. 2022. TEAM-atreides at SemEval-2022 task 11: On leveraging data augmentation and ensemble to recognize complex named entities in Bangla. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1524–1530, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971. Version 1, February 27, 2023.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

Azmine Toushik Wasi, Raima Islam, Mst Rafia Islam, Farig Sadeque, Taki Hasan Rafi, and Dong-Kyu Chae. 2025. Dialectal bias in bengali: An evaluation of multilingual large language models across cultural variations. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25)*, pages 1380–1384, Sydney, NSW, Australia. Association for Computing Machinery.

Azmine Toushik Wasi, Taki Hasan Rafi, Raima Islam, and Dong-Kyu Chae. 2024. BanglaAutoKG: Automatic Bangla knowledge graph construction with semantic neural graph filtering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2100–2106, Torino, Italia. ELRA and ICCL.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, and 17 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100. Version 4, June 27, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zheng Xin Yong, Cristina Menghini, and Stephen Bach. 2024. LexC-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13990–14009, Miami, Florida, USA. Association for Computational Linguistics.

Abdullah Khan Zehady, Shubhashis Roy Dipta, Naymul Islam, Safi Al Mamun, and Santu Karmaker. 2024. Banglallama: Llama for bangla language. *Preprint*, arXiv:2410.21200.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131. CC-BY 4.0 license.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024a. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):293:1–293:39.

Yuetong Zhao, Hongyu Cao, Xianyu Zhao, and Zhijian Ou. 2024b. An empirical study of retrieval augmented generation with chain-of-thought. In *Proceedings of the 14th IEEE International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 436–440, Singapore. IEEE.

A Dataset

A.1 Monolingual vs. Multilingual

The distinction between monolingual and multilingual datasets is crucial in low-resource language research because each provides unique insights into how LLMs learn, generalize and represent linguistic knowledge. From Table 3 we can see that the multilingual datasets such as MulD (Liu et al., 2021), mMARCO (Bonifacio et al., 2022), IndicNLG(Kumar et al., 2022a), OffMix-3L(Goswami et al., 2023), MIRACL (Zhang et al., 2023), Aya (Singh et al., 2024a), WMT24 (Deutsch et al., 2025), SeeGULL(Bhutani et al., 2024) and MultiMWP (Gamage et al., 2025) highly rely on machine translation to extend language coverage, which converts English or other high-resource languages into Bangla.

For example, the MultiCoNER (Malmasi et al., 2022) dataset, a multilingual resource for complex named entity recognition, includes 11 languages, including Bangla (BN). During dataset construction, the corpus was first developed in English using three main sources: Wikipedia (LOWNER), MS MARCO questions (MSQ NER) and Bing ORCAS search queries (ORCAS NER). English entities were automatically labeled through Wikipedia interlinks and Wikidata gazetteers. For Bangla, German, Chinese and Hindi, the English sentences were translated into these languages using the Google Translation API. Only the surrounding sentence context, excluding the entity terms, was translated. No human participation was involved in the correction process.

Although monolingual datasets do not fully rely on data translation, they cannot be considered sufficient, as they still show clear limitations in knowledge depth and cultural grounding. For example, datasets such as BanglaRQA, BanglaQuAD (Rony et al., 2024), BanNERD (Mahtab et al., 2025) and BanglaNLG (Bhattacharjee et al., 2022b) are curated by native speakers and are typically developed for task specialization and smaller size rather than broad knowledge coverage. BanglaRQA (Ekram et al., 2022) and BanglaQuAD mainly use Bangla Wikipedia as their knowledge source, while BanglaNLG draws on online forums and social media posts for generative summarization tasks. This indicates that, although the data is authentically written in Bangla, it represents only a narrow range of topics and lacks the domain diversity needed for deeper and more comprehensive language under-

Dataset	Year	Task	Sources	Size	Metrics	LM-based	Translation
BanglaRQA	2022	QA	Bangla Wiki	14.8k Pairs	F1	✓	None
BanglaHateBERT	2022	TC	Multi-domain	15k Posts	Acc., F1	✓	None
BanglaParaphrase	2022	NLU	RoarBangla	466.6k Pairs	BLEU, ROUGE-L, BERTScore	✓	Automatic
BanglaNLG	2022	NLG	Multi-domain	2.75M pairs	BLEU; ROUGE; F1	✓	Automatic
BanMANI	2023	TC	Social Media	500 Articles	F1, ROUGE-L	✓	None
M-BAD	2023	TC	Social Media	15.6k Text	F1	✓	None
BnPC	2024	NLG	Newspapers	8.7k Pairs	ROUGE-L, BERTScore	✓	None
Bengali & Banglish	2024	TC	Multi-domain	80.1k Sent.	Acc., F1, Kappa	✓	Post-edit
BNSENTMIX	2024	TC	Multi-domain	20k Sent.	Acc., F1	✓	None
BanglaQuAD	2024	QA	Bangla Wiki	30.8k Sent.	F1	✓	None
BEnQA	2024	QA	Exams Book	5.1k Sent.	F1	✓	None
BDA	2024	TC	Multi-domain	48.3k Sent.	F1, CS	✓	Automatic
BanStereoSet	2025	BE	Multi-domain	2k Sent.	SE	✓	Post-edit
ANCHOLIK-NER	2025	IE	Multi-domain	17.4k Sent.	F1	✓	Post-edit
BANTH	2025	TC	Social Media	37k Samp.	F1, Acc.	✓	Post-edit
BnMMLU	2025	IE	Multi-domain	138k Sent.	Acc., F1	✓	None
REVEAL-Bangla	2025	NLU	Wikipedia	104 Samp.	Acc.	✓	Post-edit
BanNERD	2025	IE	Multi-domain	85k Sent.	F1	✓	None
BIDWESH	2025	TC	Social Media	9.1k Sent.	Acc., F1	✓	Post-edit
BanglaMATH	2025	NLU	School Workbooks	1.7k Samp.	Acc.	✓	None
Bangla-Alpaca	2025	NLU, NLG	Multi-domain	224k Pairs	N/A	✓	Automatic
MultiLingual datasets supporting Bangla							
MulDA	2021	IE	Multi-domain	100K Sent.	F1	✓	Automatic
MultiCoNER	2022	IE	Multi-domain	26M Tokens	F1	✓	Automatic
mMARCO	2022	IR	Multi-domain	532k Pairs	Recall	✓	Automatic
IndicNLG	2022	NLG	Multi-domain	8.5M Sent.	ROUGE-L, F1	✓	Automatic
OffMix-3L	2023	TC	Social Media	100k Sent.	F1	✓	None
MIRACL	2023	IR	Wikipedia	726k Pairs	Recall	✓	None
Aya	2024	NLU, NLG	Multi-domain	513M	Acc., chrF++	✓	Automatic
WMT24	2024	MT	Multi-domain	6.1k Pairs	chrF++	✓	Post-edit
SeeGULL	2024	BE	Multi-domain	25.8k Sent.	F1	✓	Automatic
Global-MMLU	2025	NLU	Multi-domain	589.7k Sent.	Acc.	✓	Post-edit
MultiMWP	2025	NLG	Multi-domain	7.4k Sent.	F1	✓	Post-edit

Table 3: Overview of Bangla NLP datasets. "LM-based" stands for involvement of LLMs on those Dataset. "Trans." represent Translation is used dataset generation process. Symbols ✓and ✗indicate the presence or absence of those things. Abbreviations of NLP Task: QA - Question Answering; TC -Text Classification; NLG - Natural Language Generation; IE - Information Extraction; NLU - Natural Language Understanding; MT -Machine Translation; IR -Information Retrieval; BE - Bias Evaluation; SE - Stereotype Engagements.

standing.

A.2 NLP Task

Among the reviewed Bangla NLP datasets, three major task types are common: Text Classification, Language Understanding and Question Answering. Datasets such as BanglaHateBERT, BanMANI, M-BAD, BNSENTMIX, BDA, BANTH, BIDWESH and OffMix-3L were primarily developed for different forms of text classification. On the other hand, BanglaParaphrase, REVEAL-Bangla, BanglaMATH and Bangla-Alpaca were designed for language understanding, while BanglaRQA, BanglaQuAD and BEnQA focus on question-answering tasks.

NLU is a common NLP task across these datasets, covering mathematical, commonsense and general language understanding. For instance, Reveal-Bangla (Islam and Sarti, 2025) evaluates multi-step reasoning, particularly commonsense, temporal and multi-hop factual reasoning that links historical facts, performs simple calculations and applies logical inference. The reasoning follows a chain-of-thought (CoT) style, where intermediate steps lead to the final answer. While step-level importance enables analysis of model reasoning, the evaluation remains limited for non-binary examples, constraining deeper insight into complex reasoning.

IE, NLG, BE and MT tasks are less common but still represented.

A.3 Data Sources

Multi-domain is the common type of resource used in generating datasets of Bangla, covering sources such as Wikipedia, newspapers, exams, social me-

dia, online forums, and curated knowledge bases. Examples include BanglaHateBERT, BanglaNLG, BNSENTMIX, BDA, BanStereoSet, ANCHOLIK-NER, BnMMLU, BanNERD, Bangla-Alpaca etc. For instance, BanglaNLG is a benchmark created to evaluate natural language generation in Bangla across six key tasks. For machine translation, it uses the BanglaNMT parallel corpus, while the text summarization task is based on BBC News articles with editor-written summaries. The question-answering task uses the BQA dataset, which includes human-written Bangla questions and answers, along with some unanswerable ones. And those data are collected from school exam, notebooks, etc.

Wikipedia and social media are also common domains of collecting data including BanglaRQA, BanMANI, M-BAD, BanglaQuAD, BANTH, REVEAL-Bangla, BIDWESH, OffMix-3L and MIRACL. BanglaQuAD is one of the dataset which is built from Bengali Wikipedia. The authors first shortlisted 658 important Wikipedia articles across 14 categories and 114 subcategories, then had native Bengali annotators create 30,808 question answer pairs directly from those passages using their own annotation tool, BnAnno. On the other side, BIDWESH (Fayaz et al., 2025) dataset contained over 50,000 Bangla comments collected from social media platforms and online streaming sites.

Other resources are less common like Text books, Newspapers, Exam Papers etc. For example, The BEnQA (Shafayat et al., 2024a) dataset was created using official Bangladeshi national board exam questions for grades 8, 10, and 12, covering science subjects such as Mathematics, Physics, Chemistry and Biology.

A.4 Performance Metrics

The commonly used performance metric in Bangla NLP research is the F1 score. It is applied in some datasets such as BANTH, BIDWESH, BanglaHate-BERT, BDA, and M-BAD for text classification. This evaluation method calculates the standard precision (P), recall (R) and F1 score. F1 measures the overlap between system and gold tokens.

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{3}$$

Another common metric is accuracy, which measures how many total predictions are correct. Datasets like BanglaHateBERT, BDA, and BanglaMATH (Prama et al., 2025) used accuracy to summarize overall performance. For question answering datasets such as BanglaQuAD and BEnQA, the main metrics are Exact Match (EM) and token-level F1. EM measures how many answers match the gold reference exactly.

Other performance metrics include Recall@k, Mean Reciprocal Rank (MRR), BLEU, ROUGE-L, BERTScore and chrF++, which are used in different tasks. For instance, for generation and translation tasks such as BanglaParaphrase, BanglaNLG, and WMT24, the common metrics are BLEU, ROUGE-L, BERTScore and chrF++.

A.5 Translation of Data

To generate monolingual datasets, almost 60 percent of the time the technique used is translation from high-resource languages. More than half of the datasets, including BanglaParaphrase, BanglaNLG, BDA, BanStereoSet, ANCHOLIK NER, BANTH, REVEAL Bangla, BIDWESH and Bangla Alpaca, were created using translation-based methods. However, several Bangla NLP datasets were developed entirely in native Bangla without any machine or human translation, such as BanglaRQA, BanglaQuAD, BanglaHateBERT, M BAD, BanMANI, BNSENTMIX, BnPC and BanglaMATH.

Different translation strategies have been adopted across these datasets. For instance, Indic-NLG (Kumar et al., 2022a) applied parallel translation of summarization and dialogue data from English into several Indic languages, including Bangla. The Aya (Singh et al., 2024a) corpus combined automated translation models with crowd validation, mixing machine generated translations from English. The BanglaParaphrase (Akil et al., 2022) corpus used back translation, where Bangla text was first translated into English and then back into Bangla, producing paraphrased versions verified for semantic consistency.

While not all datasets used translation to create training data, some applied translation for reasoning support. For example, BANTH (Haider et al., 2025) introduced translation-based prompting, translating Banglish text into Bangla or English before inference to enhance the understanding ability of large language models. In contrast, BanglaRQA and BanglaQuAD contain manually written question and answer pairs collected from Bangla Wikipedia and news sources. Likewise, BNSENTMIX, BnPC and BanglaMATH were compiled directly from local sources such as product

reviews, newspapers and school textbooks, ensuring the natural use of the Bangla language.

A.6 Dataset Size

The size of Bangla NLP datasets varies widely, showing that data collection for this language is still developing unevenly. Generally, Bangla datasets are small, usually between 5,000 and 50,000 samples, such as BDA with 48,000, BanglaQuAD with 31,000, and BANTH with 37,000 entries. Only a few have reached large-scale levels, including BanglaNLG with 2.75 million pairs and Bangla-Alpaca with 224,000 pairs, which are important steps toward creating large, model-ready datasets. In comparison, multilingual datasets that include Bangla, such as IndicNLG with 8.5 million sentences, Aya with 513 million and Global-MMLU with about 590,000, are much larger and more balanced across languages. This clear difference in size shows that Bangla data is still limited for effective pretraining and evaluation of large language models.

B Application Areas

B.1 Fundamental NLP Tasks

From low-level preprocessing to high-level reasoning, this taxonomy (Figure 10) presents the full landscape of Bangla NLP in a single overview. All referenced studies were published between 2021 and 2025 and focus on the use of large language models. At the foundational stage, limited work exists on preprocessing tasks such as tokenization (Alam et al. 2021 Bose and Salim 2024), normalization (Fahim, 2023) and part-of-speech tagging (Nadif et al., 2024), followed by lexical and syntactic analysis that includes word sense disambiguation (Chakrabarty et al. 2021 Yong et al. 2024) and dependency or constituency parsing. For instance, the study by (Alam et al., 2021) primarily reviews core Bangla preprocessing tasks and shows that large language models can replace older rule-based and classical machine learning methods for Bangla preprocessing, provide better or similar performance and offer standard baselines and data splits for future Bangla NLP research.

B.2 Applied NLP Tasks

The Applied NLP section in this taxonomy (Figure 10) covers a wide range of topics, which reflects how active and rapidly growing this area is. It includes information retrieval, relational semantics, and both natural language understanding and natural language generation tasks. Recent research focuses on NLU (Liu et al. 2023, Wasi et al. 2024, Aurpa et al. 2023, etc.) and NLG (Fahad et al. 2024, Niloy 2024, etc), especially on text classification, question answering, information extraction, and machine translation. Maximum papers examine how large language models perform on these tasks and how LLMs can be used to improve them. For example, Hasan et al. (2025a) used LLMs in two ways for machine translation task. First, LLMs helped to generate the data, they used large LLMs like GPT-4o, Claude, and Gemini to generate translations, and humans chose the best ones. Then, the authors trained BanglaT5 (Bhattacharyya and Bhattacharya, 2025) using their new dataset.

B.3 Advancing Reasoning and Responsible NLP

In the advanced reasoning area some papers have focused on reasoning rather than responsible NLP such as ethics and bias. Different prompting-based learning methods such as CoT, zero shot, and few shot have been applied and although research in Bangla is still limited various reasoning mechanisms have been explored including multi-step reasoning (Shafayat et al., 2024b), commonsense reasoning (Ridoy et al., 2025), and inductive and deductive reasoning (Tasnim et al., 2025). In (Mohsin et al., 2025) the main reasoning style is prompting-based CoT with short stepwise explanations. For each toxic Bangla sentence the LLM is instructed to first explain in about fifteen words why the sentence is toxic then decide the class (Explicit, Implicit, or Not Toxic) and finally generate a detoxified paraphrase. Therefore the reasoning is generally local and explanation-oriented CoT rather than long multi-step logic or formal inductive and deductive proofs. When handling implicit toxicity the model also relies on simple commonsense and social reasoning.

C Approaches for Adapting LLMs

C.1 Prompting-based Adaptation

To adapt LLMs for Bangla language, different types of adaptation techniques have been used, with prompting being one of the common (Table 4). There are four main types of prompting approaches that have been applied: (1) translation-based prompting (Shafayat et al., 2024a), (2) chain-of-thought (CoT) prompting (Prama et al. 2025,

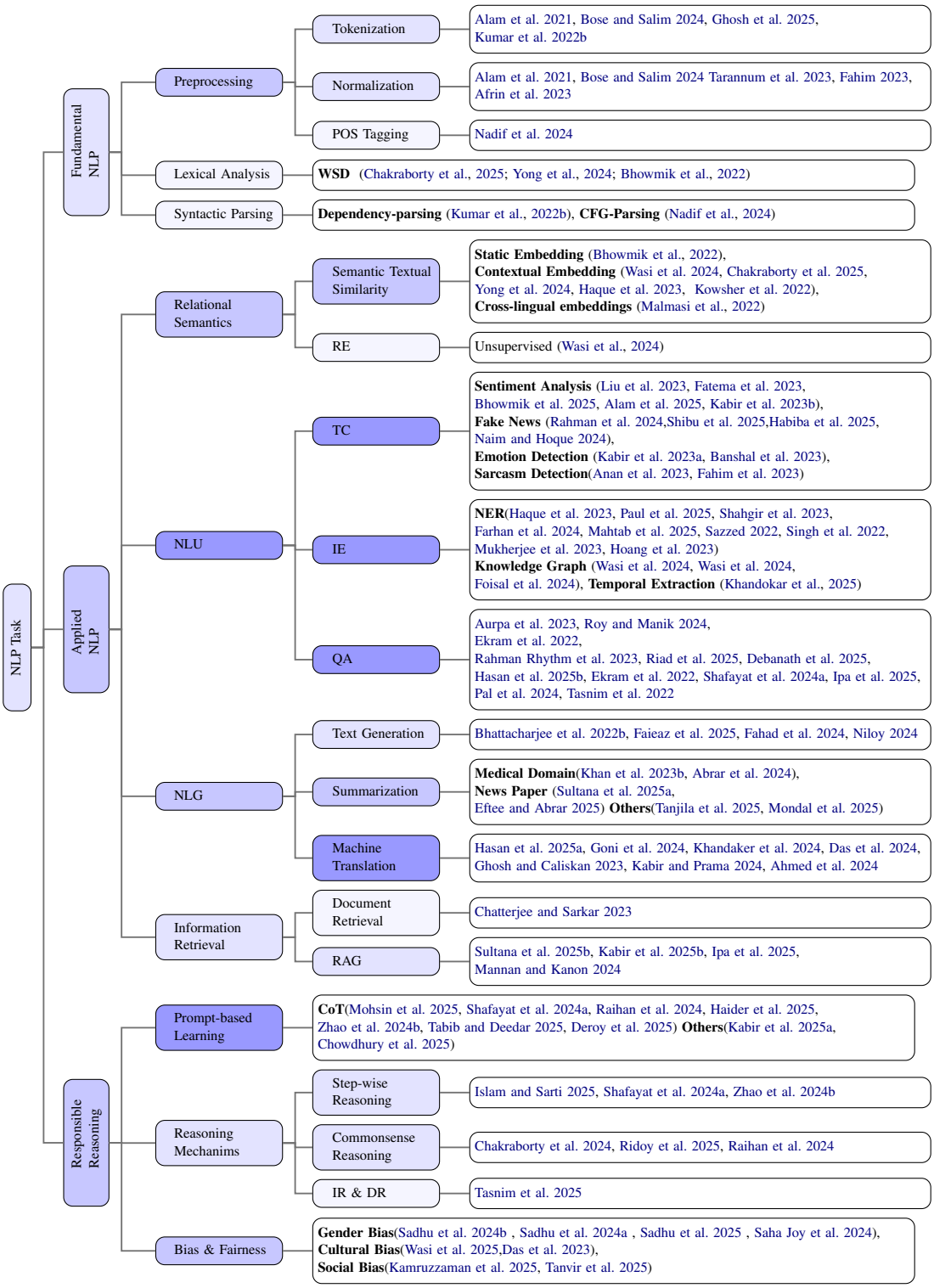

Figure 10: A complete taxonomy of Bangla language processing application areas. This taxonomy is created by taking the idea from those (Zhao et al. 2024a, Chu et al. 2024 and Wasi et al. 2024) papers. Key abbreviations: IE - Information Extraction; QA - Question Answering; MT - Machine Translation; NER - Named Entity Recognition; CoT - Chain-of-Thought; RAG - Retrieval-Augmented Generation; WSD - Word Sense Disambiguations; RE - Relation Entities; TC - Text Classification; NLU - Natural Language Understanding; NLG - Natural Language Generation; IR - Inductive Reasoning and DR -Deductive Reasoning. Color legend are : extensive research, : moderate research, : limited research, : very few works.

Paper	Prompting	Finetune	Translation	Tokenizer Adaptation	Cross-/Multilingual
Liu et al. 2021	✗	✓	✓	✗	✓
Akil et al. 2022	✗	✓	✓	✗	✓
Bhattacharjee et al. 2022b	✗	✓	✓	✓	✗
Malmasi et al. 2022	✗	✓	✓	✗	✓
Kumar et al. 2022a	✗	✓	✓	✓	✓
Goswami et al. 2023	✓	✓	✗	✗	✓
Kamruzzaman et al. 2023	✓	✓	✗	✗	✓
Monsur et al. 2023	✗	✓	✗	✓	✓
Elahi et al. 2024	✗	✓	✓	✗	✓
Singh et al. 2024b	✓	✓	✗	✗	✓
Saha et al. 2024	✓	✓	✗	✗	✗
Faisal et al. 2024b	✗	✓	✓	✗	✓
Shafayat et al. 2024a	✓	✗	✓	✗	✓
Tariquzzaman et al. 2024	✗	✓	✓	✗	✓
Singh et al. 2024a	✗	✓	✓	✗	✓
Bhutani et al. 2024	✓	✗	✓	✗	✓
Wasi et al. 2024	✓	✗	✓	✗	✓
Hasan et al. 2024	✓	✓	✗	✗	✓
Haider et al. 2025	✓	✓	✓	✗	✓
Mahtab et al. 2025	✓	✓	✗	✗	✓
Deutsch et al. 2025	✓	✗	✓	✗	✓
Singh et al. 2025	✓	✗	✓	✗	✓
Gamage et al. 2025	✗	✓	✓	✗	✓
Dewangan et al. 2025	✗	✓	✓	✗	✓
Nahin et al. 2025	✓	✓	✓	✓	✗
Dehan et al. 2025	✓	✓	✗	✗	✓
Mahfuz et al. 2025	✓	✓	✓	✓	✓

Table 4: Comparative Analysis of Bangla LLM Adaptation Strategies: Prompting, Fine-Tuning, Translation, Tokenizer Adaptation, and Cross-/Multilingual Transfer. Symbols ✓and ✗indicate the presence or absence of those things.

Shafayat et al. 2024a), (3) zero-shot prompting (Singh et al. 2025, Deutsch et al. 2025, Hasan et al. 2024, & Bhutani et al. 2024), and (4) few-shot prompting (Dehan et al. 2025, Nahin et al. 2025, Mahtab et al. 2025, Haider et al. 2025, Wasi et al. 2024, Saha et al. 2024, Kamruzzaman et al. 2023, & Goswami et al. 2023).

BEnQA (Shafayat et al., 2024a) uses two straightforward prompting ideas to help LLMs perform better in Bangla. First, it instructs the model to reason step by step in Bangla before selecting an answer, known as CoT prompting. Second, it includes the English version of the same question alongside the Bangla one, referred to as translation-appended prompting. This approach allows the model to leverage its stronger English understanding to grasp terms and concepts more effectively while still providing the reasoning and final answer in Bangla.

A few studies have also explored generative elicitation prompting to examine model bias through open-ended responses. SeeGULL Multilingual (Bhutani et al., 2024) asked language models to freely generate stereotypes so researchers could observe what the models "think".

C.2 Fine-Tuning-based Adaptation

Several studies in this review employed fine-tuning strategies to adapt LLMs Bangla language. The majority of early approaches relied on full fine-tuning, where all model parameters were updated for a specific downstream task. For instance, Liu et al. 2021, Bhattacharjee et al. 2022b, Malmasi et al. 2022, Kumar et al. 2022a, Kamruzzaman et al. 2023, Monsur et al. 2023, Elahi et al. 2024, Faisal et al. 2024b, Singh et al. 2024a & Mahtab et al. 2025.

More recent works have increasingly turned to Parameter-Efficient Fine-Tuning (PEFT) techniques to reduce training costs while retaining strong adaptation capabilities. For example Hasan et al. 2024, Mahfuz et al. 2025 & Dehan et al. 2025 used PEFT technique where LoRA adapters were attached to the linear layers of the models Dehan et al. 2025 to make faster, less memory-intensive, and cheaper, while still adapting the models to Bangla text classification tasks effectively.

Though this not very common in this study but some paper used instruction tuning to make LLMs behave more like a helpful assistant to follow human written instruction clearly. For instance, Akil

et al. (2022), Kamruzzaman et al. (2023), Tariquzzaman et al. (2024) has used Instruction tuning, where BanglaLlama (Zehady et al., 2024) was designed to make LLaMA models follow Bangla instructions, like answering questions, generating text, or solving reasoning problems in Bangla.

C.3 Tokenizer or Vocabulary Adaptation

Most LLMs are trained on English or other Latin-based languages, so they often break words into too many small pieces of tokens. This makes it harder for the model to understand or generate good text. Tokenization adaptation ((Bhattacharjee et al., 2022b), (Kumar et al., 2022a)& (Monsur et al., 2023)) solves this problem by updating the tokenizer using more Bangla text. Nahin et al. (2025) created a new Bangla-friendly tokenizer and combined it with Llama, using about 48 GB of Bangla text. They tested different vocabulary sizes so each Bangla word used fewer tokens, making the model faster and smarter. Mahfuz et al. (2025) mainly showed that English tokenizers break Bangla words too much and suggested retraining the tokenizer with more Bangla data like in BanglaT5 and NLLB.

C.4 Multilingual / Cross-Lingual Transfer

Cross-lingual transfer refers to leveraging knowledge learned from one language to enhance model performance in another language with limited resources. In this review, nearly all papers (Liu et al. 2021, Akil et al. 2022, Malmasi et al. 2022, Kumar et al. 2022a, Goswami et al. 2023, Tariquzzaman et al. 2024, Singh et al. 2024a, Bhutani et al. 2024, Hasan et al. 2024, Haider et al. 2025, Deutsch et al. 2025, Gamage et al. 2025 & Dewangan et al. 2025) applied some form of cross-lingual transfer to improve Bangla language understanding. For example, Shafayat et al. (2024a) combined Bangla and English questions within the same prompt to transfer reasoning ability from English to Bangla. Similarly, Liu et al. (2021) employed multilingual language models to generate synthetic Bangla data from English seed texts, thereby augmenting low-resource datasets and strengthening model adaptation.