Style-Aware Bangla News Headline Generation Using Large Language Models: A Comparative Zero-Shot Evaluation

Lubna Yasmin Pinky[1], Hasneen Tamanna[2], Md Ferdos kabir[3], Mohammad Ashraful Islam[2], Md. Musfique Anwar[2]
[1]Department of Computer Science and Engineering, Mawlana Bhasani Science and Technology University, Santosh, Tangail
[2]Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka-1342
[3]Department of ICT, Dhamrai Government College, Dhaka
lubnacse@mbstu.ac.bd, totinee.stu2019@juniv.edu, ashraful.islam@juniv.edu and manwar@juniv.edu
Correspondence E-mail: ashraful.islam@juniv.edu

Abstract— **News headline generation is an essential task in natural language processing (NLP), as it plays an important role in conveying the main message of a news article in a concise form. Although this task has been widely studied in English and other high-resource languages, Bangla headline generation remains less explored, especially when the headline needs to follow a specific writing style. In this study, four Large Language Models (LLMs), namely GPT-OSS-120B, LLaMA-3.3-70B, Qwen3-32B, and Qwen2.5-14B-Local, are evaluated for style-aware Bangla news headline generation in a zero-shot setting. The models were not fine-tuned for this task; instead, they were guided only through style-specific prompts. For the experiment, 75 Bangla news articles were used, and headlines were generated in six styles: Political, Informative, Dramatic, Short, Analytical, and Event-Based. A total of 1,800 generation attempts were performed. The generated headlines were evaluated using 15 automatic metrics, including BLEU-1/2/4, METEOR, ROUGE-1/2/L, WER, CER, Edit Distance, TF-IDF Cosine Similarity, Jaccard Similarity, and BERTScore Precision, Recall, and F1. The results show that LLaMA-3.3-70B achieves the best overall performance, with a BLEU-1 score of 0.1454, a METEOR score of 0.1333, and a BERTScore-F1 score of 0.2064. Qwen3-32B obtains the highest BERTScore Recall score of 0.2144, indicating better coverage of the reference headline content. The results also indicate that the Informative style produces more accurate and consistent headlines than the other styles. Overall, this study demonstrates the potential of zero-shot LLMs for style-aware Bangla headline generation, while also highlighting the limitation of using a single reference headline for evaluating creative generated outputs.**

***Index Terms*—Bangla NLP, headline generation, large language models, zero-shot learning, style-aware text generation, abstractive summarization, BERTScore, BLEU, METEOR.**

I. Introduction

Automatic news headline generation is an important task in natural language processing (NLP) and have many practical applications in journalism, information retrieval, and content summarization. A well-written headline must capture the main point of article in brief and engaging way. For Bengali (Bangla) language, which is spoken by more than 230 million people worldwide and is sixth most spoken language in world, automatic headline generation remains challenging and underexplored research area [1].

The emergence of large language model (LLM) such as GPT series [2], LLaMA family [3], and Qwen series [4] have open new possibilities for text generation task in both high-resource and low-resource language. Large language models can follow human instructions and generate text in different styles or formats via natural languages. However, systematic evaluation of these model for Bangla headline generation with multiple style conditioning have not been done in prior work.

Style-conditioned text generation refer to the task of generating text that follow specific stylistic guideline or tone [5]. In the context of news headlines, different styles may be appropriate for different audiences or purposes:

i) **Political:** Focus on government actions and policy implications.
ii) **Informative:** Prioritize factual accuracy and completeness.
iii) **Dramatic:** Use emotional language and rhetorical devices.
iv) **Short:** Sacrifice detail for brevity (≤8 words).
v) **Analytical:** Provide interpretive framing and causality.
vi) **EventBased:** Emphasize the specific event that occurred.

GPT, LLaMA, Qwen, and similar models can follow instructions to generate text in a specific style or format. This ability makes them useful for style-conditioned generation tasks. Motivated by this, the present study explores how well large language models can generate Bangla news headlines in different writing styles without task-specific fine-tuning. In this study, each model is given the same news content along with a specific style instruction, and the generated headline is then collected for evaluation. The paper compares the performance of selected LLMs across different headline styles and news categories using automatic evaluation metrics. Through this evaluation, the study aims to identify which models perform better for Bangla headline generation, which styles produce more reliable outputs, and what limitations remain when creative Bangla headlines are evaluated using a single reference headline.

II. Literature Review

A. Bangla Natural Language Processing

Bangla natural language processing has received growing attention in recent years, but it is still less-resourced

compared to English and many other high-resource languages. Early Bangla NLP research mainly focused on basic linguistic tasks such as morphological analysis and part-of-speech tagging. Alam et al. [6] worked on morphological analysis and POS tagging of Bangla words, which is an important early contribution to computational processing of Bangla text.

With the development of transformer-based models, Bangla NLP research has moved toward benchmark construction and pre-trained language models. Alam et al. [7] examined Bangla NLP benchmarks and highlighted the issue of lexical overlap, showing the need for careful evaluation. Bhattacharjee et al. [8] introduced BanglaBERT, a pre-trained language model for Bangla language understanding, which achieved strong performance on several downstream tasks.

For Bangla summarization and headline-related generation, research is still limited. Hasan et al. [9] introduced XL-Sum, a large-scale multilingual abstractive summarization dataset covering 44 languages, including Bangla. Their work provides an important resource for Bangla summarization. In addition, large language models have opened new possibilities for Bangla text generation. Brown et al. [10] showed that large language models can perform different tasks through prompting, which is relevant for zero-shot headline generation without task-specific fine-tuning.

B. Automatic Evaluation Metrics for Text Generation

Automatic evaluation of generated text is a major challenge in NLP. BLEU, proposed by Papineni et al. [11], measures n-gram overlap between generated and reference text and is widely used in machine translation and text generation. However, BLEU has limitations, especially when several valid outputs can express the same meaning. Callison-Burch et al. [12] showed that BLEU may not always correlate well with human judgment, which is important for creative tasks such as headline generation.

METEOR [13] addresses some limitations of BLEU by considering unigram matching with both precision and recall. ROUGE [14], originally designed for summarization evaluation, measures recall-oriented overlap and is also useful for headline evaluation because a headline can be treated as a short summary. However, these lexical metrics may fail to capture semantic similarity when generated headlines use different wording from the reference.

Neural evaluation metrics are increasingly used to capture meaning beyond surface word overlap. BERTScore [15][16] computes similarity using contextual embeddings and can better handle paraphrasing and semantic variation. This is useful for Bangla, where morphological variation and flexible wording may reduce lexical overlap. BLEURT [17] is another learned evaluation metric based on BERT, although such metrics may require suitable language-specific adaptation for low-resource languages.

C. Style-Conditioned Text Generation

Style-conditioned text generation aims to generate text according to a specific style, attribute, or writing requirement. Earlier work in style transfer used neural methods to separate style from content. Shen et al. [18] proposed style transfer from non-parallel text using cross-alignment, while Hu et al. [5] reviewed major approaches, challenges, and evaluation methods in text style transfer.

With large language models, style control can be performed more directly through prompting. Keskar et al. [19] introduced CTRL, where control codes guide text generation with specific attributes. Wei et al. [20] showed that instruction-tuned models can follow task instructions in a zero-shot setting. This supports the idea that GPT, LLaMA, Qwen, and similar models can be prompted to generate Bangla headlines in different styles without fine-tuning.

In news headline generation, style is important because different audiences and purposes may require different headline forms. An informative headline may focus on clarity and completeness, while a dramatic headline may use more expressive language. However, style-conditioned headline generation is still underexplored for Bangla. Moreover, style instructions may not transfer equally across languages due to differences in journalistic tone, culture, and expression. Ahuja et al. [21] emphasized multilingual evaluation of generative AI, which is relevant for assessing LLMs in Bangla generation. Therefore, a systematic evaluation of style-aware zero-shot Bangla headline generation is needed.

D. Research Gap

Although headline generation and style-conditioned text generation have been widely studied in English, Bangla headline generation is still less explored. Existing Bangla NLP studies mostly focus on benchmarks, classification, language understanding, or general summarization, but not on style-aware headline generation. Moreover, multilingual summarization work does not clearly show how different LLMs perform when Bangla headlines are generated with specific style instructions. Therefore, this study addresses this gap by systematically evaluating four LLMs for style-aware zero-shot Bangla news headline generation across different styles and categories.

III. Dataset Description

A. Data Source

To prepare the evaluation we used the publicly available dataset Bangladesh News Article from Kaggle [22]. It contains Bangla news articles from major online news portals in Bangladesh. The articles covered different topics, such as politics, sports, entertainment, technology, international news, and economy. From this collection, we selected 75 articles to create a balanced sample for the experiment. Each selected article was paired with its original human-written headline, which was used as the reference headline for evaluation.

B. Experimental Subset Statistics

Table I summarize the overview of the 75 selected news article used in this study.

Table I: Experimental Dataset Statistics

Statistic	Value
Total articles selected	75
Headline styles per article	6
Models evaluated	4
Total generation attempts	1,800
Valid generations	1,698 (94.3%)
Missing/NaN (GPT-OSS-120B)	102 (5.7%)
Avg. article length (tokens)	312.4
Avg. reference headline (chars)	47.2

C. Reference Headline Characteristics

The reference headlines were obtained directly from the source news portals, ensuring they represent professional, publication-ready examples authored by journalists. These headlines follow a brief, telegraphic style, with an average length of 47.2 characters and 7.8 words. Furthermore, every reference headline follows standard written Bangla, avoiding the use of colloquialisms or regional dialects to maintain linguistic consistency across the dataset.

D. Category Coverage

The 75 articles are distributed across six topical domains: Politics (18 articles, 24%), Sports (14 articles, 18.7%), Entertainment (12 articles, 16%), Technology (11 articles, 14.7%), International (10 articles, 13.3%), and Business/Economy (10 articles, 13.3%).

IV. Methodology

This study follows a style-aware zero-shot evaluation framework for Bangla news headline generation. The overall workflow includes style-specific prompt design, zero-shot generation using API-based and local LLMs, automatic evaluation, and comparative analysis. The complete workflow of the proposed methodology is shown in Figure 1.

A. Style Definitions

Six headline styles are used for style-conditioned generation, as defined in Table II. Each style represents a different way of presenting the main news event or message. Although many other headline styles are possible, such as question-based, persuasive, or neutral styles, this study focuses on six commonly useful styles for Bangla news headline generation: Political, Informative, Dramatic, Short, Analytical, and Event-Based.

Table II: Headline Style Definitions

Style	Definition
Political	Emphasizes political actors, power dynamics, and governance implications
Informative	Neutral, factual; answers who/what/where/when
Dramatic	Emotionally charged; uses strong verbs and creates urgency
Short	Concise (≤8 words); telegraphic style
Analytical	Provides context, causality, or implication beyond the surface event
EventBased	Centers on the specific event or action as the main subject

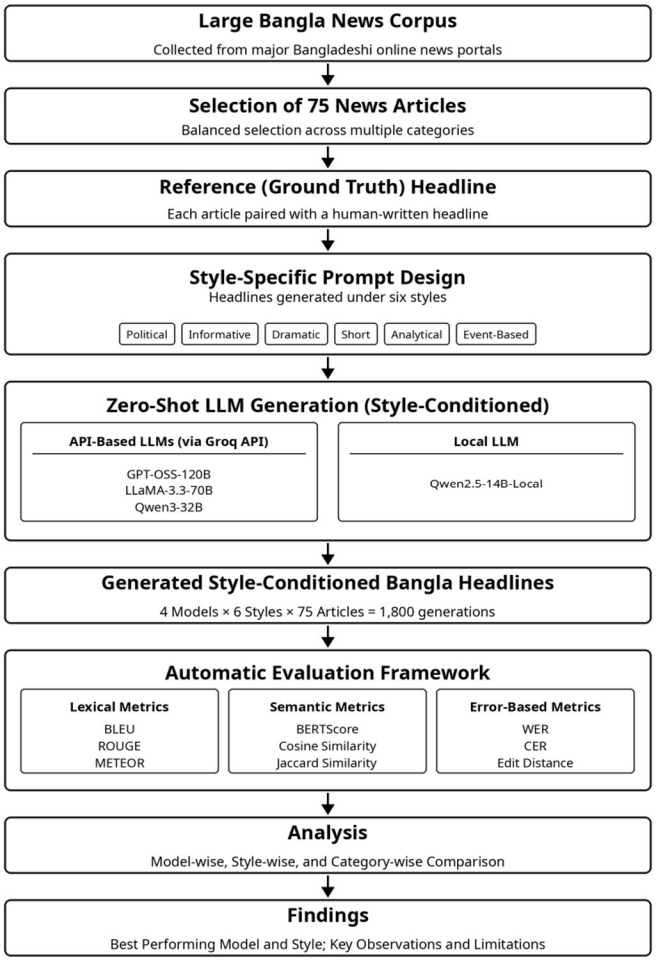

Fig. 1. Flowchart of Bangla headline generation process.

B. Large Language Models

This study evaluates four large language models for style-aware zero-shot Bangla news headline generation: GPT-OSS-120B, LLaMA-3.3-70B, Qwen3-32B, and Qwen2.5-14B-Local. These models were selected to cover different model scales, context capacities, and deployment conditions. The selected models include very large API-accessed models as well as a locally deployable quantized model, which allows the study to observe both high-capacity and practical local generation settings.

GPT-OSS-120B is the largest model used in this study and belongs to the open-weight GPT-OSS model family. LLaMA-3.3-70B is an instruction-tuned model from Meta's LLaMA family and is designed for multilingual text-based generation.

Qwen3-32B is a multilingual instruction-following model from the Qwen family, while Qwen2.5-14B-Local represents a smaller locally deployable model used in quantized format. Table III summarizes the main technical characteristics of the selected LLMs.

Table III: Technical overview of the large language models used in this study

Model	Parameters	Context Window (tokens)	Model Type
GPT-OSS-120B	117B total; 5.1B active	131,072	Open-weight MoE reasoning model
LLaMA-3.3-70B	70B	128K / 131,072	Instruction-tuned multilingual LLM
Qwen3-32B	32.8B	32,768	Multilingual instruction-following LLM
Qwen2.5-14B-Local	14.7B	131,072	Locally deployable instruction model

C. Zero-Shot Headline Generation Process

In the headline generation stage, each model received the news article content together with one style instruction. For every article, the model generated one headline for each of the six styles. This process was repeated for all four models. As a result, the complete experiment involved 4 models, 6 styles, and 75 articles, producing a total of 1,800 generation attempts.

D. Evaluation Metrics and Equations

Different evaluation metrics are applied to evaluate the result.

a) N-gram Overlap Metrics

BLEU-n measures modified n-gram precision with a brevity penalty:

$$BLEU\text{-}n = BP \cdot exp\left(\sum_{k=1}^{n} w_k \, log p_k\right)$$

where p_k is the modified k-gram precision, $w_k = 1/n$, and $\text{BP} = \min(1, e^{1-r/c})$ with r the reference length and c the candidate length.

METEOR extends precision/recall with stemming and alignment:

$$METEOR = F_{mean} \cdot (1 - Pen)$$

where $F_{mean} = \frac{10P_uR_u}{R_u+9P_u}$ is the harmonic mean of unigram precision P_u and recall R_u, and $Pen = 0.5\left(\frac{chunks}{unigrams\_matched}\right)^3$ penalizes fragmentation.

b) Recall-Oriented Overlap Metrics

ROUGE-N computes n-gram recall:

$$ROUGE\text{-}N = \frac{\sum_{s\in\mathcal{R}} \sum_{g_n\in s} Count_{match}(g_n)}{\sum_{s\in\mathcal{R}} \sum_{g_n\in s} Count(g_n)}$$

ROUGE-L uses the longest common subsequence (LCS):

$$ROUGE\text{-}L = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

c) Error-Based Metrics

WER and **CER** measure edit operations normalized by reference length:

$$WER = \frac{S+D+I}{N_{ref}}, \quad CER = \frac{S_c+D_c+I_c}{N_c}$$

where S, D, I denote substitutions, deletions, and insertions at word (WER) or character (CER) level.

d) Semantic Similarity Metrics

TF-IDF Cosine Similarity measures vector-space alignment:

$$Cosine(h,r) = \frac{h \cdot r}{\| h \| \| r \|}$$

Jaccard Similarity measures token-set overlap:

$$J(H,R) = \frac{|H \cap R|}{|H \cup R|}$$

BERTScore computes contextual embedding similarity:

$$BERTScore\text{-}F1 = 2 \cdot \frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

V. Experimental Setup

A. Hardware and System Configuration

For models with large parameters, we used API-based inference (GPT-OSS-120B, LLaMA-3.3-70B, Qwen3-32B) and experiment was conducted via the Groq Cloud platform. For local inference, Qwen2.5-14B-Local was executed on the researcher's Windows-based personal computer using LM Studio. The system was equipped with an NVIDIA GeForce RTX 3080 Ti Laptop GPU with 16 GB GDDR6 VRAM and 32 GB RAM. The model was used in quantized GGUF format, allowing the 14B-parameter model to run within the available local GPU memory.

B. Prompt Template

fixed prompt template was used for all models to maintain consistency in the generation process. The same template was applied across all articles, styles, and models, where only the style name and article text were changed.

```
You are a Bangla news headline writer.
Given the following Bangla news article,
write a {style} style headline in Bangla.
```

```
The headline should be in Bangla language only.
Keep the headline concise (under 15 words).
Article: {article_text}
Write only the headline, nothing else:
```

C. Checkpoint Mechanism

To ensure fault tolerance during large-scale generation, a file-based checkpoint system is implemented. After each successful generation, the result is immediately written to disk. On restart, already-completed article–style–model combinations are detected and skipped, preventing redundant API calls.

VI. Results and Analysis

A. Overall Model Performance

Table IV summarizes the overall performance of the four models across all evaluation metrics. LLaMA-3.3-70B shows the strongest result in lexical-similarity based metrics including BLEU-1: 0.1454, METEOR: 0.1333, Cosine: 0.2715, Jaccard: 0.1081. In contrast, Qwen3-32B performs best in semantic recall, achieving the highest BERTScore-R score of 0.2144. However, LLaMA-3.3-70B achieves the best overall semantic alignment, with the highest BERTScore-F1 score of 0.2064, while Qwen3-32B obtains a BERTScore-F1 score of 0.1922. GPT-OSS-120B records the lowest scores for most automatic metrics. However, manual observation suggests that GPT-OSS-120B often produced more creative and longer headlines.

The model-wise comparison of BLEU-1, BLEU-2, and BLEU-4 scores is shown in Figure 2.

Table IV: Overall Model Performance on Core Metrics

Metric	GPT	LLaMA	Qwen3	Qwen2.5
Parameters	117B total 5.1B active	**70B**	**32B**	**14B**
BLEU-1	0.0813	**0.1454**	0.1224	0.0932
BLEU-2	0.0393	**0.0799**	0.0601	0.0390
BLEU-4	0.0175	**0.0357**	0.0275	0.0183
METEOR	0.0741	**0.1333**	0.1099	0.0853
Cosine	0.1824	**0.2715**	0.2562	0.2226
Jaccard	0.0561	**0.1081**	0.0869	0.0644
BS-F1	0.1367	**0.2064**	0.1922	0.1579
BS-Recall	0.1520	0.2128	**0.2144**	0.1920
WER	**1.1796**	1.3285	1.4402	1.6981

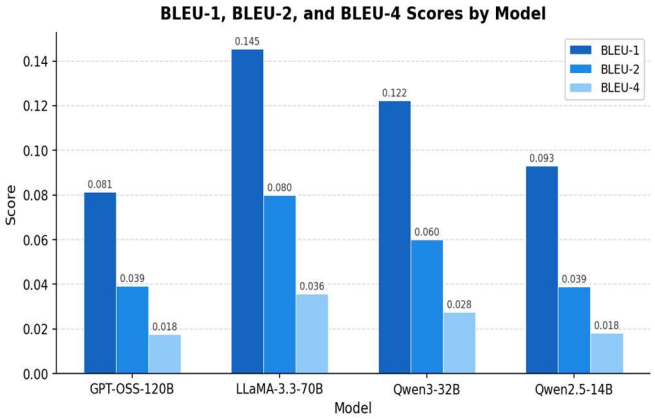

Fig. 2. BLEU-1, BLEU-2, and BLEU-4 scores by model.

Result indicates, LLaMA records the highest scores across all BLEU variants. The model wise comparison of METEOR and BERTScore-F1 scores is shown in Figure 3. It shows LLaMA-3.3-70B performs best on METEOR; whereas Qwen3-32B is competitive on BERTScore-F1.

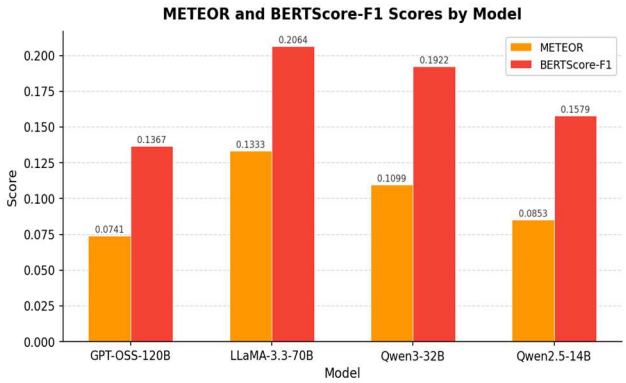

Fig. 3. METEOR and BERTScore-F1 by model.

B. BERTScore Detailed Results

Table V represents the detailed BERTScore. LLaMA-3.3-70B performs best in terms of precision (0.2191) and F1 (0.2064), while Qwen3-32B stands out in recall (0.2144), indicating better coverage of reference content. In contrast, GPT-OSS-120B obtains lowest scores on all BERTScore measures.

Table V: BERTScore Precision, Recall, and F1 by Model

Model	BS-P	BS-R	BS-F1
GPT-OSS-120B	0.1326	0.1520	0.1367
LLaMA-3.3-70B	**0.2191**	0.2128	**0.2064**
Qwen3-32B	0.1866	**0.2144**	0.1922
Qwen2.5-14B-Local	0.1431	0.1920	0.1579

The BERTScore Precision, Recall, and F1 comparison across models is shown in Figure 4. We can see, LLaMA-3.3-70B dominates in precision; Qwen3-32B leads in recall.

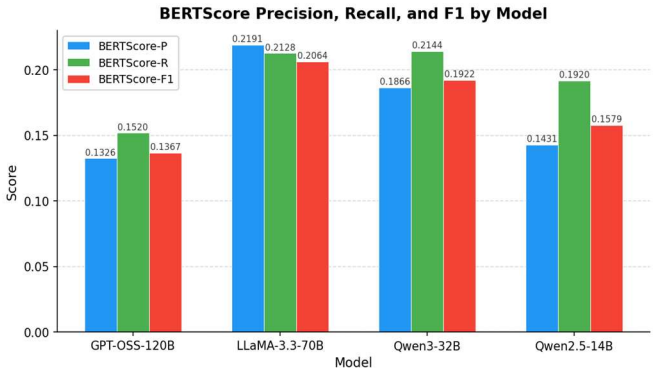

Fig. 4. BERTScore Precision, Recall, and F1 by model.

C. *Per-Style Performance*

Table VI presents metric scores assembled by headline style. The Informative style consistently shows the best lexical alignment (BLEU-1: 0.1229, METEOR: 0.1196, BERTScore-F1: 0.1862). The Short style achieves the lowest WER (1.0436) and the highest BERTScore-Precision (0.1968) most likely due to its concise nature. Contrary to this, the Dramatic style shows the most creative deviation, resulting in the lowest BLEU-1 (0.0998).

Table VI: Performance by Headline Style (All Models Combined)

Style	BLEU-1	METEOR	Cosine	BS-F1	WER
Informative	**0.1229**	**0.1196**	**0.2521**	**0.1862**	1.4320
EventBased	0.1165	0.1093	0.2394	0.1792	1.4928
Analytical	0.1088	0.1032	0.2372	0.1743	1.5546
Short	0.1100	0.0863	0.2199	0.1643	**1.0436**
Political	0.1056	0.0931	0.2346	0.1772	1.4501
Dramatic	0.0998	0.0925	0.2161	0.1589	1.4959

The style-wise comparison of BLEU-1 and METEOR scores is shown in Figure 5.

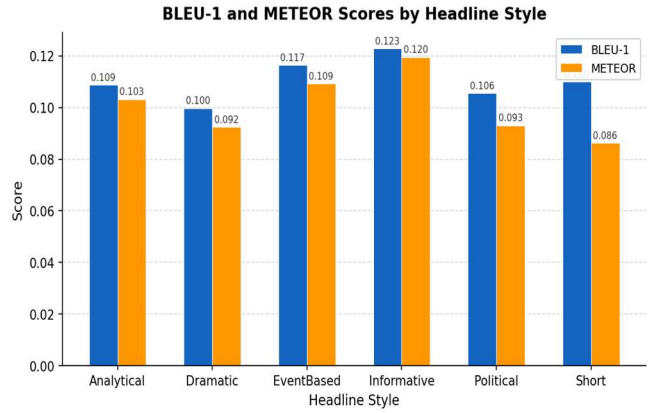

Fig. 5. BLEU-1 and METEOR by headline style (all models combined).

The Informative shows the highest scores, whereas the Dramatic style shows the lowest scores.

D. *Model × Style Interaction: BLEU-1*

Table VII reports BLEU-1 scores across all model × style combinations. LLaMA-3.3-70B achieves the highest BLEU-1 score in the Informative style (0.1632). While GPT-OSS-120B consistently performs the worst across all styles, largely due to missing predictions.

Table VII: Model × Style BLEU-1 Scores

Style	GPT	LLaMA	Qwen3	Qwen2.5
Analytical	0.0752	**0.1541**	0.1339	0.0720
Dramatic	0.0699	**0.1398**	0.1189	0.0706
EventBased	0.0843	**0.1558**	0.1378	0.0881
Informative	0.0891	**0.1632**	0.1451	0.0942
Political	0.0761	**0.1492**	0.1219	0.0752
Short	0.0932	**0.1099**	0.1128	0.1341

E. *Model × Style Interaction: METEOR*

Table VIII shows METEOR scores for each model × style combinations. LLaMA-3.3-70B achieves the highest scores in both Informative (0.1521) and EventBased (0.1489) styles. Qwen2.5-14B-Local delivers strong results in the Short style (0.1187), suggesting that quantized local models can compete with API-served models when generating concise outputs.

Table VIII: Model × Style METEOR Scores

Style	GPT	LLaMA	Qwen3	Qwen2.5
Analytical	0.0681	**0.1452**	0.1278	0.0717
Dramatic	0.0623	**0.1289**	0.1143	0.0645
EventBased	0.0789	**0.1489**	0.1354	0.0740
Informative	0.0842	**0.1521**	0.1389	0.0832
Political	0.0643	**0.1298**	0.1143	0.0640
Short	0.0868	0.0899	0.0887	**0.1187**

F. *Model × Style Interaction: BERTScore-F1*

Table IX presents BERTScore-F1 results for all model × style combinations. Qwen3-32B performs best in BERTScore-F1 in Informative (0.2134) and Analytical (0.2089) styles, while LLaMA-3.3-70B takes the lead in EventBased (0.2198) and Political (0.2156) headlines.

Table IX: Model × Style BERTScore-F1 Scores

Style	GPT	LLaMA	Qwen3	Qwen2.5
Analytical	0.1423	0.2067	**0.2089**	0.1393
Dramatic	0.1312	0.1934	0.1878	0.1234
EventBased	0.1489	**0.2198**	0.2143	0.1334
Informative	0.1567	0.2098	**0.2134**	0.1489
Political	0.1398	**0.2156**	0.2089	0.1445
Short	0.1267	0.1872	0.1701	0.1632

G. Headline Length Analysis

The model-wise comparison of WER and CER values is shown in Figure 6.
Qwen2.5-14B-Local shows the highest error rates; GPT-OSS-120B achieves the lowest WER.

To observe model and style-wise headline generation, Table X presents selected candidate headlines for the reference headline "জুলাই সনদ নিয়ে বিতর্ক শেষ হলো না কেন?" along with their evaluation scores.

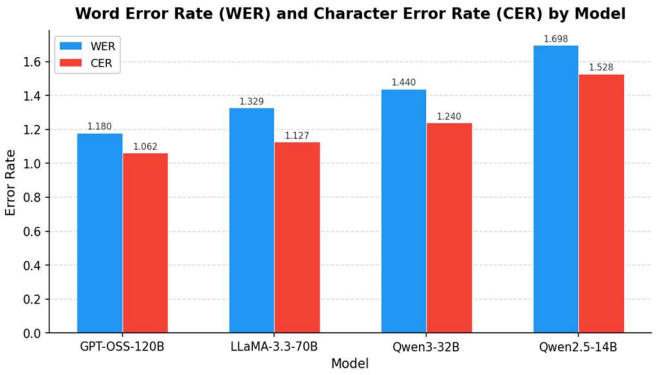

Fig. 6. WER and CER by model.

Table X: Sample Metric-Based Comparison of Candidate Headlines with a Reference Headline

Style	Model	Generated Headlines	BLEU_1	BLEU_2	BLEU_4	METEOR	WER	CER	Cosine Similarity	Jaccard Similarity
Analytical	*gpt-oss-120b*	আট-মাসের আলোচনার পর জুলাই চুক্তি স্বাক্ষর, তবে এনসিপি-এর অংশ না থাকা ও আহতদের সংশোধনীতে তীব্র বিতর্ক অব্যাহত	0.18	0.03	0.02	0.17	1.88	2.41	0.17	0.14
	llama-3.3-70b	জুলাই সনদের বিতর্ক: এনসিপির অংশগ্রহণ নিয়ে উত্তেজনা	0.12	0.04	0.03	0.06	0.88	**0.84**	**0.27**	0.07
	qwen2.5-14b-local	জুলাই সনদের মোকাবিলায় এনসিপি-অভ্যুত্থানের দল: আলোচনা অব্যাহত, শেষ সংশোধনীও বিতর্কিত	0.20	0.05	0.03	0.12	1.13	1.62	0.24	0.13
	qwen3-32b	জুলাই সনদ স্বাক্ষরের পরেও বিরূপ বিতর্কের অব্যাহত থাকা	**0.25**	**0.19**	**0.06**	**0.23**	**0.75**	0.92	0.24	**0.14**
Dramatic	*gpt-oss-120b*	আট মাসের তিক্ত আলোচনার পরেও জুলাই সনদে জ্বলে ওঠে বিতর্কের শিখা!	0.09	0.03	0.02	0.06	1.38	1.35	0.22	0.06
	llama-3.3-70b	আটমাস আলাপ-আলোচনার পরেও শেষ হয়নি বিতর্ক	0.24	0.06	0.03	0.13	**0.88**	**0.81**	0.20	0.17
	qwen2.5-14b-local	জুলাই আন্দোলনের শেষ মুহূর্তে ধমক না সমঝোতা: এনসিপি ছাড়িয়া বিতর্ক অবাধ	0.36	0.06	0.03	0.24	1.25	1.35	0.28	0.27
	qwen3-32b	জুলাই সনদ স্বাক্ষরের পরেও আগুনে পুড়ছে রাজনীতি, বিতর্ক শেষ হলো না!	**0.45**	**0.37**	**0.12**	**0.58**	**0.88**	1.08	**0.39**	**0.36**
Event Based	*gpt-oss-120b*	আট মাসের আলোচনার পর জুলাই সনদ স্বাক্ষর, তবে বিতর্ক অব্যাহত রইল	0.27	0.17	0.04	0.31	1.25	1.24	0.24	0.19
	llama-3.3-70b	জুলাই সনদ স্বাক্ষর, থেমে থাকার নাম বিতর্ক	**0.37**	**0.23**	**0.07**	**0.32**	**0.75**	**0.68**	**0.33**	**0.25**
	qwen2.5-14b-local	"জুলাই সনদের অনুমোদন পর্যন্তও বিতর্ক চলছে: এনসিপি-এর অংশগ্রহণ থেকে আহতদের স্বাস্থ্য পর্যন্ত নানা বিষয়ে মোহর চিকন	0.06	0.02	0.01	0.06	1.88	2.30	0.19	0.04
	qwen3-32b	জুলাই সনদ স্বাক্ষরের পরেও বিরোধিতা শেষ	**0.37**	**0.23**	**0.07**	**0.32**	**0.75**	0.70	0.27	**0.25**

Style	Model	Generated Headlines	BLEU_1	BLEU_2	BLEU_4	METEOR	WER	CER	Cosine Similarity	Jaccard Similarity
		হয়নি								
Informative	*gpt-oss-120b*	আট মাসের আলোচনার পরও জুলাই সনদের স্বাক্ষরে বাকি বিতর্ক	0.22	0.05	0.03	0.12	1.13	1.22	0.23	0.13
	llama-3.3-70b	জুলাই সনদ স্বাক্ষর: শেষ মুহূর্তের বিতর্ক এড়ানো গেল না কেন?	**0.60**	**0.37**	**0.07**	**0.62**	**0.63**	**0.81**	**0.48**	**0.50**
	qwen2.5-14b-local	জুলাই সনদ স্বাক্ষরের পরও: আলোচনায় অবাধ্য নানা বিতর্ক, এনসিপির অপার্থিবতা ও আহতদের স্বার্থ	0.15	0.11	0.03	0.22	1.38	1.81	0.22	0.11
	qwen3-32b	জুলাই সনদ স্বাক্ষরের পর বিতর্ক শেষ হয়নি, প্রশ্ন উঠে শেষ মুহূর্তের সংশোধনী এবং এনসিপির অংশগ্রহণ	0.27	0.20	0.04	0.36	1.38	1.78	0.27	0.22
Political	*gpt-oss-120b*	জুলাই সনদ স্বাক্ষরিত হলেও বিতর্কে গর্জে উঠল, এনসিপি বাদে নেতৃত্বের ঝুঁকি বাড়াল!	0.17	0.12	0.04	0.22	1.25	1.43	0.25	0.11
	llama-3.3-70b	জুলাই সনদ স্বাক্ষর: বিতর্কের অন্ত নেই	0.24	**0.19**	**0.06**	**0.24**	**0.75**	**0.57**	**0.32**	0.17
	qwen2.5-14b-local	জুলাই সনদের অবশেষে স্বাক্ষর: আহতদের বিষয়ে শেষ মুহূর্তের বিতর্ক এবং নেতা পক্ষগোলয়ন্ডি	**0.25**	0.05	0.02	0.18	1.25	1.59	0.27	**0.18**
	qwen3-32b	জুলাই সনদের পরেও বিরোধ বন্ধ হয়নি	0.12	0.04	0.03	0.06	0.88	0.59	0.20	0.08
Short	*gpt-oss-120b*	আট মাসের আলোচনার পরও জুলাই সনদের বিতর্ক অব্যাহত	0.25	0.06	0.03	0.13	1.00	1.08	0.23	0.14
	llama-3.3-70b	জুলাই সনদ স্বাক্ষর, অব্যাহত বিতর্ক	0.33	**0.21**	**0.07**	**0.33**	0.75	0.65	0.31	**0.30**
	qwen2.5-14b-local	জুলাই সনদে অক্টোপাল আলোচনা, বিতর্ক অবিরত	0.24	0.06	0.03	0.13	0.88	0.73	0.29	0.17
	qwen3-32b	জুলাই সনদে বিতর্ক নিরাপদে মুক্তি না পাওয়া	**0.37**	0.07	0.04	0.19	**0.63**	**0.62**	**0.37**	0.25

As shown in Table X, Qwen3-32B generates the headline that is closest to the reference in terms of lexical overlap, achieving the highest BLEU-1, BLEU-2, BLEU-4, and METEOR scores with a comparatively lower WER. On the other hand, LLaMA-3.3-70B obtains the highest cosine similarity, which suggests stronger semantic closeness even when the exact word overlap is lower. Overall, the table shows that model performance varies across different evaluation metrics. Therefore, both lexical and semantic metrics are necessary for a fair evaluation of generated Bangla headlines.

The model-wise comparison of TF-IDF Cosine Similarity and Jaccard Similarity is shown in Figure 7. The figure suggest, LLaMA-3.3-70B achieves the highest TF-IDF Cosine Similarity and Jaccard Similarity scores, with values of 0.2715 and 0.1081. Qwen3-32B follows closely, while GPT-OSS-120B obtains the lowest scores. This indicates that LLaMA-3.3-70B generated headlines with stronger surface-level semantic similarity and word overlap with the reference headlines.

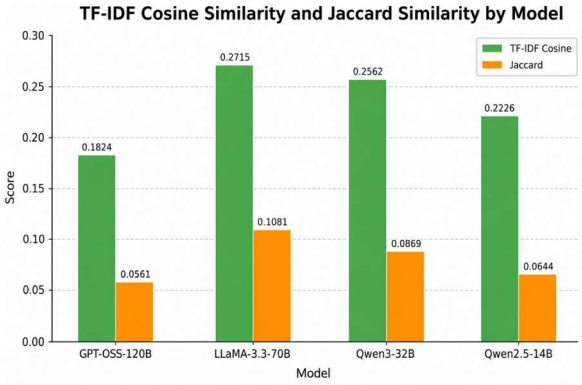

Fig. 7. TF-IDF Cosine and Jaccard Similarity by model.

The overall core metric comparison across all evaluated models is shown in Figure 8. Although GPT-OSS-120B performs lower in most automatic metrics, manual observation suggests that it sometimes generates longer and more creative headlines. These outputs may differ from the reference wording and therefore receive lower automatic scores, even when they remain meaningful.

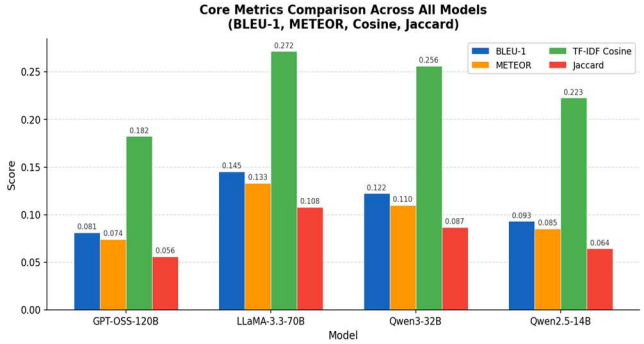

Fig. 8. Core metrics comparison across all models.

The figure show that LLaMA-3.3-70B performs best in BLEU-1, METEOR, TF-IDF Cosine Similarity, Jaccard Similarity, and BERTScore-F1. Qwen3-32B ranks second and shows strong semantic recall. GPT-OSS-120B and Qwen2.5-14B-Local perform lower in most metrics, although GPT-OSS-120B

The style-wise comparison of BERTScore-F1 is shown in Figure 9. The Informative style achieves the strongest semantic alignment with the reference headlines, indicating that factual and complete headline instructions help the models generate more reliable outputs. Other styles, such as Dramatic and Short, show lower BERTScore-F1 because they often change wording, tone, or detail level, which reduces similarity with the single reference headline.

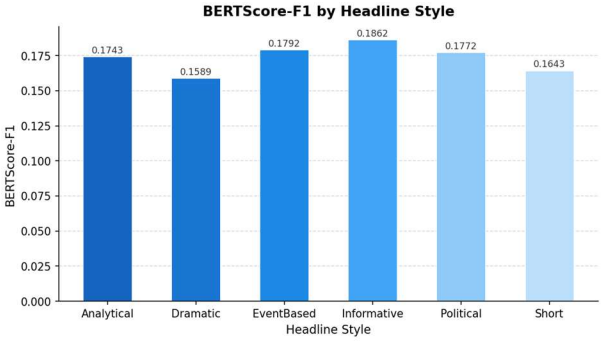

Fig. 9. BERTScore-F1 by headline style.

Informative headlines achieved the highest semantic alignment; while Dramatic presented the lowest.

H. Conclusion

A. Key Findings

(i) **LLaMA-3.3-70B performs best overall**, achieving the highest BLEU-1 score of **0.1454**, METEOR score of **0.1333**, and BERTScore-F1 score of **0.2064**. Based on these results, it can be considered the most reliable model for Bangla headline generation in this study. Basically LLaMA-3.3-70B model follows the headline instruction more consistently and generates concise outputs that remain close to the reference headline. Most automatic metrics reward similarity in wording, structure, and meaning. Therefore, LLaMA benefits by producing headlines that are both meaningful and closer to the human-written reference.

(ii) **Qwen3-32B performs best in semantic recall**, achieving the highest BERTScore-R score of **0.2144**. This suggests that it captures more reference-related content, although its lexical overlap is lower than that of LLaMA-3.3-70B. Qwen3-32B shows strong semantic recall, meaning that it often captures key information from the article, although its wording may differ slightly from the reference headline.

(iii) **Qwen2.5-14B-Local shows promising performance in the local deployment setting**, especially considering its smaller parameter size. This indicates that locally deployed LLMs can be useful for cost-aware and privacy-aware Bangla headline generation.

(iv) **GPT-OSS-120B obtains lower scores in most automatic metrics** because some of its outputs were missing, while many of its valid outputs were more creative and longer. These creative headlines may still be relevant and readable, but they may use different words, add extra framing, or follow a different expression style than the single reference headline. As a result, metrics such as BLEU, METEOR, Jaccard similarity, WER, and edit distance may penalize these outputs.

(v) **The Informative style produces the most consistent and reference-aligned headlines**, while the Dramatic style shows greater creative variation. This suggests that factual and complete style instructions are more effective for automatic Bangla headline generation.

(vi) **Automatic metrics have limitations for Bangla headline evaluation**, especially when only one reference headline is used. A low overlap score does not always mean that a generated headline is poor, because several valid headlines can be written for the same news article.

B. Limitations and Future Works

This study has several limitations. First, it uses a single reference headline for each article, which may not capture other valid paraphrases or stylistic variations. Second, the dataset is limited in size, so the findings may not fully represent all Bangla news domains. Third, the study relies on automatic metrics, which may not completely measure

fluency, readability, factual correctness, or cultural appropriateness.

Future work should include multiple reference headlines, larger Bangla news corpora, human evaluation, Bangla-specific fine-tuning, and retrieval-augmented generation for better factual grounding.

ACKNOWLEDGEMENT

This study has been done under the research project "Generating Bangla News Headlines Using Large Language Models" sponsored by the Faculty of Mathematical and Physical Sciences, Jahangirnagar University. Our research interests in social media analysis, namely friend recommendation and event detection, were also a significant motivation for this work, in which short textual content, user interaction and event-oriented information are important. This prompted us to look into the potential of using LLM to create a succinct and meaningful Bangla news headline from a full news article. AI (Grammarly and Chat GPT) have only been used to enhance grammar, linguistic clarity and the visual representation of the block diagram.

REFERENCES

[1] JEthnologue. "Bengali: A language of Bangladesh." SIL International, 2023. [Online]. Available: https://www.ethnologue.com/language/ben.

[2] T. B. Brown et al., "Language models are few-shot learners," in *Proc. NeurIPS*, 2020, pp. 1877–1901.

[3] H. Touvron et al., "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[4] J. Bai et al., "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[5] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proc. ICML*, 2017, pp. 1587–1596.

[6] F. Alam et al., "BanglaBERT: Combating embedding barrier for low-resource language understanding," in *Proc. NAACL-HLT (Findings)*, 2022, pp. 2695–2706.

[7] T. Hasan et al., "XL-Sum: Large-scale multilingual abstractive summarization for 44 languages," in *Proc. ACL-IJCNLP (Findings)*, 2021, pp. 4693–4703.

[8] A. Bhattacharjee et al., "BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource NLG in Bangla," *arXiv:2205.11081*, 2023.

[9] S. A. Chowdhury et al., "Towards Bangla automatic text summarization," in *Proc. ECIR*, 2021, pp. 1–12.

[10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

[11] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop*, 2005, pp. 65–72.

[12] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summ. Branches Out*, 2004, pp. 74–81.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. ICLR*, 2020.

[14] M. Post, "A call for clarity in reporting BLEU scores," in *Proc. WMT*, 2018, pp. 186–191.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[16] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[17] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[18] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[19] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," in *Proc. NeurIPS*, 2017, pp. 6830–6841.

[20] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A conditional transformer language model for controllable generation," *arXiv preprint arXiv:1909.05858*, 2019.

[21] J. Wei et al., "Finetuned language models are zero-shot learners," in *Proc. ICLR*, 2022.

[22] Haque Siam, E. (2026). *Bangladesh News Articles Dataset* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/9384161

Biography

Lubna Yasmin Pinky is an Associate Professor in the Department of Computer Science and Engineering at Mawlana Bhashani Science and Technology University, Bangladesh. She received her B.Sc. and M.Sc. degrees in Computer Science and Engineering from Jahangirnagar University. She is currently pursuing an MPhil, focusing on social media data analysis using artificial intelligence and natural language processing techniques. Her research interests include natural language processing, large language models, deep learning, social media analytics, text classification, and AI-assisted education.

Hasneen Tamanna completed her Bachelor of Science (Eng.) in Computer Science and Engineering from Jahangirnagar University, Bangladesh, where she is currently pursuing her Master of Science (Eng.) degree. Her research interests include Natural Language Processing (NLP), Machine Learning, and Artificial Intelligence, with a particular focus on developing computational solutions for low-resource languages. She is enthusiastic about applying intelligent technologies to tackle meaningful real-world challenges.

Md. Ferdos Kabir is a Lecturer in ICT and a member of the BCS General Education Cadre. He holds BSc (Honor's) and MSc degrees in CSE from Jahangirnagar University, Bangladesh. His research interests include artificial intelligence, machine learning, educational technology, digital pedagogy, information systems, and ICT-enabled teaching and learning. His work particularly focuses on the application of emerging technologies to improve the quality, accessibility and effectiveness of education.

Mohammad Ashraful Islam received his B.Sc. and M.Sc. degrees in Computer Science and Engineering from Jahangirnagar University, Bangladesh. He is currently serving as a faculty member in the Department of Computer Science and Engineering. His research interests include artificial intelligence, machine learning, natural language processing, social media data analysis, image processing, and computational intelligence. He has published several research articles in national and international journals, conference proceedings and a book on computational intelligence.

Md Musfique Anwar has awarded PhD degree from Swinburne University of Technology, Melbourne, Australia in 2018. He has received M.Sc. degree from Kyoto University, Japan in 2013 and B.Sc. degree in Computer Science and Engineering from Jahangirnagar University, Bangladesh in 2006. Since 2008, he is a faculty member having current designation Professor in the Department of Computer Science and Engineering of Jahangirnagar University. Currently his research focuses on Data Mining, Social Network Analysis, Natural Language Processing and Software Engineering.