A Comprehensive Text Optimization Approach to Bangla Summarization

Irtifa Haider
Department of Computer
Science and Engineering
Jahangirnagar University,
Bangladesh
irtifa.stu2019@juniv.edu

Shanjida Alam
Department of Computer
Science and Engineering
Jahangirnagar University,
Bangladesh
shanjida.stu2019@juniv.edu

Md. Tazel Hossan
Department of Computer
Science and Engineering
Jahangirnagar University,
Bangladesh
tazel.stu2017@juniv.edu

Md. Musfique Anwar
Department of Computer
Science and Engineering
Jahangirnagar University,
Bangladesh
manwar@juniv.edu

Tanjim Taharat Aurpa
Department of Data
Science and Engineering
University of Frontier
Technology, Bangladesh
aurpa0001@uftb.ac.bd

Abstract

The task of Bengali text optimization demands not only the generation of concise and coherent summaries but also grammatical accuracy, semantic appropriateness, and factual reliability. This study presents a dual-phase optimization framework for Bengali text summarization that integrates entity-preserving preprocessing and abstractive generation with mT5, followed by refinement through sentence ranking, entity consistency enforcement, and optimization with instruction-tuned LLMs such as mBART. Evaluations using ROUGE, BLEU, BERTScore, and human ratings of fluency, adequacy, coherence, and readability show consistent gains over baseline summarizers. By embedding grammatical and factual safeguards into the summarization pipeline, this study establishes a robust and scalable benchmark for Bengali NLP, advancing text optimization research. Our model achieves 0.54 ROUGE-1 and 0.88 BERTScore on BANS-Data, outperforming recent multilingual baselines.

1 Introduction

In an era of information overload, producing clear, concise, and contextually relevant text is crucial for effective communication. *Text optimization*, refining summaries for readability, semantic accuracy, and purpose alignment, is increasingly essential. This includes enforcing grammatical soundness and precise meaning for diverse audiences in the digital age.

For Bangla, a morphologically rich language with a large speaker base yet limited NLP resources, optimization addresses unique challenges such as complex grammar (e.g., case markers, verb conjugations) to deliver high-quality, inclusive content. Despite its widespread use, Bangla lags behind high-resource languages in research and tooling, necessitating specialized models and pipelines (Haque et al., 2020).

This study tackles these issues by creating a two-phase pipeline that processes Bangla text, corrects errors (e.g., missing verbs, excessive conjunctions), and evaluates summaries with tailored metrics. Advances in Bangla NLP include large-scale resources such as a 100M-word corpus used for high-accuracy spell/grammar checking (Hossain et al., 2021) and XL-Sum's multilingual benchmark with Bangla coverage (Hasan et al., 2021).

Bangla's under-representation highlights the need for integrated optimization, unlike English's advanced LLM ecosystems, especially for domains such as news and healthcare. Early surveys such as Haque et al. (2020) catalog 14 Bangla summarization methods and emphasize resource gaps. Rahman et al. (2024) improve extractive ranking with Word2Vec embeddings but do not address grammar or factuality. Hasib et al. (2023) couple an extractive BenSumm stage with (Bangla)T5 for news, yet lack explicit faithfulness optimization. Miazee et al. (2025) explore neural abstractive pipelines but face scalability limits due to small datasets. On the grammar side, Hossain et al. (2024) (*Panini*) outperform BanglaT5 on a 7.7M-pair GEC corpus, while Sultana et al. (2024) show LLM potential for Bangla news summarization alongside cross-lingual/cultural gaps. Rule-based grammar pattern detection (Prapty et al., 2021) complements these efforts.

Therefore, we introduce a two-phase hybrid summarization–optimization model: **Phase 1** performs preprocessing, NER, BERT-style embed-

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 151–160
December 23, 2025 ©2025 Association for Computational Linguistics

dings, and uses mT5 to draft summaries; **Phase 2** optimizes with entity-aware ranking, redundancy reduction, and mBART-based refinement to improve readability, reduce hallucination, and preserve factual consistency.

We then combine automatic and human assessments: ROUGE (recall), BLEU (precision), and BERTScore (semantic similarity), plus human judgments on fluency, adequacy, coherence, and faithfulness. This research, based on the implemented pipeline, pursues the following objectives:

- Text Optimization via Grammar Correction (e.g., missing verbs, excessive conjunctions) and ensuring entity accuracy; assess with ROUGE, BLEU, and BERTScore.
- Two-Phase Hybrid Summarization Model combining classical NLP (tokenization, NER, stemming) with transformers (BERT embeddings, mT5 drafting, mBART refinement) to enhance readability and reduce redundancy.
- Comprehensive Evaluation using automated metrics (ROUGE, BLEU, BERTScore) and human evaluations (fluency, adequacy, coherence, faithfulness) to address prior Bangla ATS limitations.

By filling gaps in integrated text optimization, this work paves the way for future research in Bangla NLP, with potential to transform automated digital content for Bangla-speaking communities worldwide.

2 Related Work

The task of Automatic Text Summarization (ATS) for Bangla involves generating concise and coherent summaries from textual or structured data inputs. Research has evolved from traditional extractive methods toward optimization-driven and neural approaches, with growing interest in multi-document inputs (Haque et al., 2020; Wahab et al., 2024). Despite solid progress in Bangla NLP and optimization techniques, challenges persist due to limited datasets, linguistic complexity, and computational constraints (Haque et al., 2020; Hasan et al., 2021).

Several works have developed frameworks, surveys, and benchmarks tailored to Bangla, underscoring the importance of resource creation and evaluation (Haque et al., 2020; Hossain et al., 2021; Hasan et al., 2023). While progress is visible, Bangla remains under-represented in large-scale multilingual corpora and lacks mature optimization-driven pipelines—clear priorities for future work (Hasan et al., 2021; Wahab et al., 2024).

Early Bangla efforts emphasized extractive techniques for their simplicity and modest resource demand. Haque et al. (2020) survey 14 Bangla summarization methods and outline evaluation hurdles and resource scarcity. Rahman et al. (2024) compare frequency/rule-based/Word2Vec extractive models and show embeddings better capture semantics, though their scope remains ranking-only without grammar or factual checks. Foysal et al. (2021) introduce Bangla-ExtraSum, comparing five extractive approaches (incl. transformer-based and Word2Vec-assisted models) on 200 prior + 500 new articles with dual human references; they report strong scores (e.g., F1 ≈ 0.68/0.63 on two sets) but stay strictly extractive. Khan et al. (2023) present a hybrid extractive model combining NER, keywords, POS, and sentence cues over a curated 960-passage, 8-domain dataset (2,880 human summaries), achieving competitive precision/F1 yet still within extractive, traditional metrics.

To address multi-document needs, Hasan et al. (2023) release BUSUM-BNLP, a 1,000-article update-summarization corpus across six domains with human references; simple TF-IDF baselines outperform SentenceRank there, but the work remains extractive and ranking-focused.

With neural architectures, Bangla abstraction gains traction. Miazee et al. (2025) propose an abstractive pipeline with neural modeling and preprocessing; data scale and evaluation remain modest. Hayat et al. (2023) benchmark multiple transformer variants (incl. fine-tuned BanglaT5) and report best ROUGE-2 ≈ 13.83 for BanglaT5. Hasib et al. (2023) build a hybrid pipeline where an extractive BenSumm stage feeds (Bangla)T5; the hybrid beats direct T5 on news, though evaluation is mostly ROUGE/BLEU. Beyond news, Barsha and Uddin (2023) compare BanglaT5 vs. pointer-generator networks for Tagore short-story summarization; feature-augmented pointer-generator wins on ROUGE, highlighting the value of linguistically informed architectures for literary text.

Foundational resources for correctness run in parallel. Hossain et al. (2021) develop large-scale Bangla spell/grammar checking (100M-word corpus; 1M-word lexicon) with strong accuracy but outside summarization. Prapty et al. (2021) apply CFG+CYK for Bangla grammar pattern detection, effective yet rule-dependent. Hossain et al. (2024) introduce *Panini*, a transformer-based GEC trained on a 7.7M parallel corpus, outperforming BanglaT5; while not a summarizer, such GEC can be integrated into summarization post-editing for fluency and correctness.

Optimization methods underpin ATS from heuristic selection to meta-heuristics. Wahab et al. (2024) review optimization-driven ATS (e.g., GA, PSO), stressing accuracy–cost trade-offs and real-time constraints. Classical probabilistic approaches (e.g., Naïve Bayes + topic words) illustrate optimization flavors in single-syllable languages (Thu, 2014), though mostly handcrafted and extractive. At the training level, optimizer choice (Adam, RMSProp, etc.) materially affects ROUGE and convergence for abstractive news summarization (Kumari et al., 2023).

Broader multilingual/cross-domain work informs Bangla directions. Hasan et al. (2021) release XL-Sum (44 languages incl. Bangla), a key benchmark, though Bangla remains relatively small/noisy. Semantic generalization with deep models improves rare-word handling in English settings (Kouris et al., 2019). Domain-specific Bangla evaluations with LLMs emerge in health (Abrar et al., 2024) and question generation (Faieaz et al., 2025), while multimodal chart-to-text for Bangla is newly explored in Chart-Summ (Tanjila et al., 2025). For Bangla news, comparative studies of LM/LLMs indicate potential but also gaps in faithfulness and robustness (Sultana et al., 2024). Overall, recent LLM-based summarization highlights controllability and scale, yet Bangla still needs optimization-guided, resource-aware adaptation and evaluation beyond ROUGE (e.g., factuality, hallucination, and grammar checks) (Wahab et al., 2024; Hasib et al., 2023; Hayat et al., 2023). Recent multilingual models such as BanglaT5, Qwen2.5-Instruct, and LLaMA-3-Instruct demonstrate promising zero-shot summarization capabilities.

3 Methodology

This study proposes a two-phase framework for Bengali text summarization: summarization and optimization (Figure 1). The pipeline combines classical NLP with transformer-based and instruction-tuned LLMs. Preprocessing removes noise (stopwords, URLs, digits), while a Named Entity Recognition (NER) module preserves key entities.

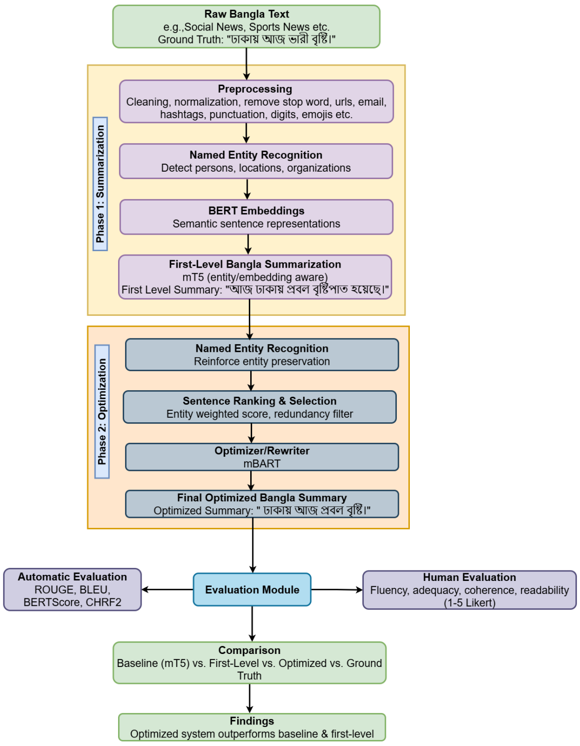
Raw Bangla Text
e.g.,Social News, Sports News etc.
Ground Truth: "ঢাকায় আজ ভারী বৃষ্টি।"
Phase 1: Summarization
Preprocessing
Cleaning, normalization, remove stop word, urls, email, hashtags, punctuation, digits, emojis etc.
Named Entity Recognition
Detect persons, locations, organizations
BERT Embeddings
Semantic sentence representations
First-Level Bangla Summarization
mT5 (entity/embedding aware)
First Level Summary: "আজ ঢাকায় প্রবল বৃষ্টিপাত হয়েছে।"
Phase 2: Optimization
Named Entity Recognition
Reinforce entity preservation
Sentence Ranking & Selection
Entity weighted score, redundancy filter
Optimizer/Rewriter
mBART
Final Optimized Bangla Summary
Optimized Summary: " ঢাকায় আজ প্রবল বৃষ্টি।"
Automatic Evaluation
ROUGE, BLEU, BERTScore, CHRF2
Evaluation Module
Human Evaluation
Fluency, adequacy, coherence, readability (1-5 Likert)
Comparison
Baseline (mT5) vs. First-Level vs. Optimized vs. Ground Truth
Findings
Optimized system outperforms baseline & first-level

Figure 1: Text Optimization Framework for Bengali Text

3.1 Summarization Techniques

Bengali summarization methods fall into two main categories: extractive, which select important sentences from the source, and abstractive, which generate new sentences that condense meaning. Early work relied on rule-based heuristics (e.g., sentence position, keywords) or machine learning models using features such as TF–IDF and entity frequency. While these methods are efficient, they often miss deeper semantics. Abstractive approaches, on the other hand, leverage neural architectures. Initial RNN+attention models struggled with long sequences, but transformer-based models like mT5 and mBART have since improved fluency and semantic coverage, producing more natural summaries.

3.2 Optimization with LLMs

Recent advances incorporate instruction-tuned large language models (LLMs) to refine draft summaries into coherent, concise, and entity-faithful outputs. In our work, we employ mBART for intermediate optimization. There are also larger models such as LLaMA-3.1-8B-Instruct and Qwen2.5-7B-Instruct for refining draft summaries while preserving entities and handling complex structures, removing redundancy, and producing concise outputs under strict length constraints.

3.3 Phase - 1 : Summarization

3.3.1 Dataset and Preprocessing

We use the Bengali Abstractive News Summarization Dataset (BANSData) (Bhattacharjee et al., 2021), containing 19,096 articles and corresponding human-written summaries across domains such as politics, economy, sports, and culture. Articles range from 5–76 words, with summaries of 3–12 words, making the dataset well-suited for abstractive summarization. A preprocessing pipeline was applied to clean and normalize the corpus by removing URLs, digits, hashtags, and redundant punctuation. Tokenization was performed using BNLP and NLTK tokenizers, stopwords were removed via the BengaliCorpus list, and BanglaStemmer was used for morphological normalization. To ensure factual accuracy, a Named Entity Recognition (NER) module preserved critical entities (persons, locations, organizations) throughout the pipeline, yielding an entity-aware dataset for embedding, summarization, and optimization.

3.3.2 Named Entity Recognition (NER)

NER is central to the framework for preserving factual reliability in low-resource Bengali summarization. Using BNLP's BengaliNER toolkit, entities such as "বাংলাদেশ ব্যাংক" (Bangladesh Bank) and "ঢাকা" (Dhaka) are detected and explicitly protected from stemming or removal. During optimization, an additional verification step ensures that all entities present in the source are retained in the final summary. This integration minimizes semantic drift and enhances the trustworthiness of generated outputs.

3.3.3 Sentence Embeddings

To capture semantic depth, each preprocessed sentence is encoded into dense vector representations using BERT embeddings, which model both left and right contexts. These embeddings serve three purposes: (i) reduce redundancy by detecting semantically overlapping sentences, (ii) strengthen ranking by promoting diverse content coverage, and (iii) enrich abstractive summarization with contextualized features that complement entity preservation. Combined with NER, this dual-layered approach ensures factual reliability, coherence, and a strong foundation for high-quality Bengali summaries.

3.3.4 Summarization with mT5

The first stage employs the mT5 model, a multilingual sequence-to-sequence transformer pretrained on diverse languages, including Bengali. Leveraging embeddings and preserved entities, mT5 generates draft summaries that are fluent and contextually meaningful. For instance, it can condense "জাপানের প্রধানমন্ত্রী শিনজো আবের সরকার নতুন করে একটি বাজেট অনুমোদন করেছে... ("Japan's Prime Minister Shinzo Abe's government has approved a new budget...") into shorter, coherent forms. However, drafts may still contain redundancy, factual drift, or incomplete coverage, requiring refinement in Phase 2.

3.4 Phase - 2 : Optimization

3.4.1 Sentence Ranking and Selection

To refine mT5 drafts, a ranking module filters sentences based on heuristic criteria. Sentences containing named entities are prioritized through an entity weight, while a length score penalizes those that are overly short or verbose. Redundancy is reduced using BERT embeddings to remove semantic overlaps, and semantic diversity is encouraged to ensure broad coverage of different aspects. Together, these steps yield summaries that are concise, entity-aware, and semantically rich, forming the basis for final optimization.

3.4.2 Entity Preservation Enforcement

Before final rewriting, BNLP's BengaliNER toolkit verifies entity consistency (e.g., "বাংলাদেশ ব্যাংক" (Bangladesh Bank), "ঢাকা" (Dhaka)). Missing or altered entities are corrected, preventing factual distortion and ensuring reliable summaries.

3.5 Optimizer/Rewriter (mBART)

The last stage uses mBART, a multilingual sequence-to-sequence transformer pre-trained with denoising objectives. Given the source article and entity-preserved draft, it:

- Enhances readability with fluent Bengali,
- Removes residual redundancy,
- Preserves named entities, and
- Improves semantic alignment with the source.

3.6 Evaluation Metrics

Evaluation of system-generated summaries is performed using both automatic metrics and human judgment. To ensure robust assessment, we adopt widely recognized metrics: ROUGE, BLEU, and BERTScore. These metrics collectively capture lexical overlap, n-gram precision, semantic similarity, and character-level fluency, providing a comprehensive evaluation of both raw summaries and optimized summaries.

3.6.1 Automatic Evaluation Metrics

To ensure robust assessment, we employ these complementary evaluation metrics:

- **ROUGE-N** measures recall by evaluating n-gram overlaps (e.g., unigrams, bigrams) between system and reference summaries.
- **BLEU** measures precision of n-gram matches, originally for machine translation, with a brevity penalty to avoid overly short outputs.
- **BERTScore** computes semantic similarity using contextual embeddings (e.g., BERT, BanglaBERT) and cosine similarity between tokens.

Final optimized summaries, generated by the LLM refinement stage (mBART), were assessed using these four automatic metrics.

3.7 Human Evaluation

Three native Bengali linguists (each with $\geq$ 3 years of experience in NLP or linguistics) independently rated a stratified random sample of 200 summaries (balanced across domains) on a 5-point Likert scale for fluency, adequacy, coherence, and readability. The raters were blinded to system identity, and ties were allowed. Inter-annotator agreement was measured using Krippendorff's α (ordinal) for each dimension; disagreements were resolved by majority vote. We report per-dimension means ± 95% CIs and α. Where:

- Fluency: grammatical correctness and sentence naturalness.
- Adequacy: extent to which the draft captured essential content.
- Coherence: logical progression and organization.
- Readability: overall ease of comprehension.

The evaluation strategy was designed to assess the effectiveness of both phases of the proposed framework. Specifically, human evaluation was applied after Phase 1 (Summarization) to assess draft summaries generated by mT5 and refined by sentence ranking, while automatic metrics were applied after Phase 2 (Optimization) to benchmark the final optimized summaries against human-written references.

4 Experimental Setup and Evaluation

The framework was meticulously tested on the Bengali Abstractive News Summarization Dataset (BANSData), a robust corpus comprising 19,096 article-summary pairs, utilizing models such as mT5 and mBART. The dataset was initially split into four parts, which underwent iterative refinement. Performance was systematically assessed using a suite of automatic metrics—ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and BERTScore—alongside detailed human evaluations focusing on fluency, adequacy, coherence, and readability, scored on a 1–5 Likert scale by a panel of three expert native Bangla linguists with linguistic experience.

4.1 Experimental Setup

Dataset. We employ the Bengali Abstractive News Summarization Dataset (BANSData), a curated collection of Bengali news articles and their corresponding human-written summaries obtained from `bangla.bdnews24.com`. The corpus comprises 19,096 article–summary pairs across diverse domains such as politics, economy, sports, culture, and social issues, and has been widely used in prior Bangla summarization research. Key hyperparameters for both phases are summarized in Table 1. Settings follow standard Hugging Face Transformer defaults, with mixed-precision training on a single NVIDIA A100 GPU.

Baseline Model. The baseline system is based on the sequence-to-sequence Long Short-Term Memory (LSTM) network with attention proposed by (Bhattacharjee et al., 2021). The model employs an encoder–decoder architecture with local attention to generate fluent and human-like abstractive summaries from Bangla news articles. The authors also released the Bengali Abstractive News Summarization (BANS) dataset, the largest publicly available Bangla summarization corpus, comprising over 19,000

Parameter	**mT5 (Phase 1)**	**mBART (Phase 2)**
Optimizer	AdamW (β_1=0.9, β_2=0.999, ϵ=1 × 10^{-8})	AdamW (same)
Learning rate	3×10^{-4}	1×10^{-5}
Warm-up / Decay	Linear (5%) / Linear	Linear (same)
Weight decay	0.01	0.01
Label smoothing	0.1	0.1
Batch size	8 × 8 (accum.=64)	8 × 8 (accum.=64)
Epochs (max)	5 (early stop patience 3)	2
Steps (approx.)	∼35k	∼10k
Token limits	512 (input), 100 (output)	512 (input), 100 (output)
GPU	A100 (40 GB)	A100 (40 GB)
Seed	42	42
Dataset split	80 / 10 / 10	80 / 10 / 10

Table 1: Fine-tuning hyperparameters for mT5 (Phase 1) and mBART (Phase 2). Both configurations use linear warm-up (5%), linear decay, and fixed seed = 42. Experiments were conducted on BANSData (train/dev/test = 80/10/10) using a single A100 GPU.

news articles and human-written summaries collected from bangla.bdnews24.com and published on Kaggle. Their system demonstrated competitive performance in both quantitative metrics (BLEU, ROUGE) and human evaluation, establishing a strong benchmark for BANSData. For comparison, this work adopts csebuetnlp/mT5_multilingual_XLSum, a multilingual T5 variant fine-tuned for abstractive summarization. The model is configured with a maximum input length of 512 tokens, maximum output length of 100 tokens, a minimum output of 30 tokens, a six-beam search, and a no-repeat 3-gram constraint to reduce redundancy. Summaries are generated using a standardized pipeline comprising tokenization, abstractive generation, and post-processing, with outputs stored in CSV format for subsequent evaluation.

Evaluation. Performance was assessed using ROUGE-1, ROUGE-L, BLEU, and BERTScore. In addition, three native Bengali linguists provided human ratings on fluency, adequacy, coherence, and readability using a five-point Likert scale. This combination of automatic and human evaluation established a robust baseline for the subsequent optimization experiments.

4.2 Automated Metric Results

Figure 2 presents the quantitative evaluation results in terms of ROUGE-1, ROUGE-L, BLEU, and BERTScore across the baseline, first-level, and optimized summaries. The baseline system yielded modest scores (ROUGE-1: 0.30, ROUGE-L: 0.31, BLEU: 0.30, BERTScore: 0.52). First-level summaries demonstrated consistent improvements, with notable gains in BLEU (0.42) and BERTScore (0.55), alongside moderate increases in ROUGE metrics. The optimized summaries achieved the best performance overall, reaching ROUGE-1: 0.54, ROUGE-L: 0.36, BLEU: 0.50, and BERTScore: 0.88. These findings indicate that progressive optimization not only improves lexical overlap but also substantially enhances semantic fidelity, with BERTScore showing the largest relative gain. Recent Bangla and multilingual summarization baselines were also evaluated for comparison, as summarized in Table 2.

Model (Configuration)	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU**	**BERTScore (F1)**
BanglaT5 (fine-tuned; Hayat et al. 2023)	0.41	0.14	0.39	0.32	0.82
Qwen2.5-7B-Instruct (0-shot)	0.44	0.16	0.40	0.35	0.84
LLaMA-3-Instruct (0-shot)	0.43	0.15	0.39	0.34	0.83
mT5 (Phase 1; ours)	**0.46**	**0.17**	**0.33**	**0.42**	**0.55**
mT5 → mBART (Phase 2; ours)	**0.54**	**0.20**	**0.36**	**0.50**	**0.88**

Table 2: Comparison on BANSData between the proposed dual-phase optimization framework and recent Bangla/multilingual baselines.

As shown in Table 2, the proposed framework delivers substantial gains over both Bangla-specific and multilingual baselines. Compared with BanglaT5, Phase 2 yields +0.13 ROUGE-1 and +0.18 BLEU improvements, while surpassing instruction-tuned LLaMA-3 and Qwen 2.5 models across all metrics. These results highlight the effectiveness of the two-phase optimization strategy in achieving superior lexical precision, semantic consistency, and factual reliability within Bengali abstractive summarization.

4.3 Human Evaluation Results

Figure 3 reports average human ratings (1 to 5 Likert scale) across four dimensions: fluency, ade-

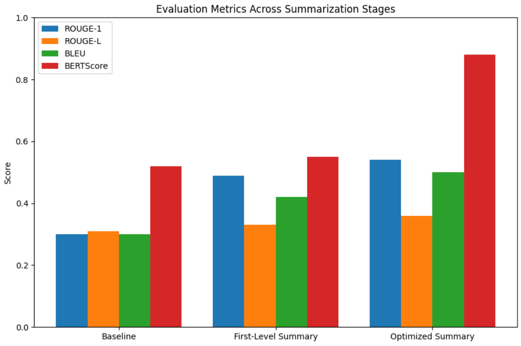

Figure 2: Automated metric results

quacy, coherence, and readability. The first-level summaries received mid-scale ratings between 3.8 to 4.1 on average, indicating reasonable but imperfect quality. Optimized summaries consistently outperformed them, scoring 4.3 to 4.5 across all dimensions. Gains were most notable in fluency and readability, suggesting that the optimization stage effectively improved grammatical correctness, sentence naturalness, and ease of comprehension. Adequacy and coherence also showed consistent improvements, highlighting the benefit of entity preservation and redundancy reduction in the final optimization phase.

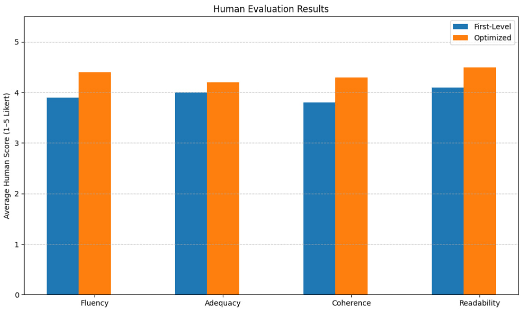

Figure 3: Human evaluation scores across fluency, adequacy, coherence, and readability

The Figure 4 (radar chart) visualizes the ratings for fluency, coherence, content preservation, and relevance across different models. The proposed model outperforms all baseline models, especially in content preservation and relevance, which are critical for generating useful and informative summaries.

Taken together, human and automatic evaluations demonstrate that the proposed two-phase framework yields consistent improvements over both the baseline and first-level drafts. Human judges valued the readability and fluency gains, while automatic scores highlighted advances in semantic alignment and lexical coverage. This dual evidence supports the effectiveness of combining entity-aware draft generation with LLM-based optimization for Bangla summarization.

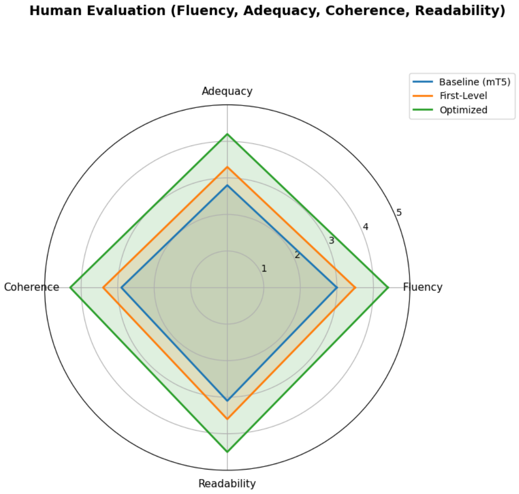

Figure 4: Radar chart of human evaluation scores (1–5) for Fluency, Adequacy, Coherence, and Readability across Baseline (mT5), First-Level, and Optimized models

4.4 Text Sample Comparison

To illustrate the qualitative impact of the proposed framework, sample outputs from the three stages of summarization are compared against the human-written reference summary. The examples highlight how the system evolved from the baseline through the first-level summarizer to the final optimized summary. To better understand the behavior of each stage, a stratified sample of 50 summaries from the baseline, Phase 1 (mT5), and the final optimized system was manually examined. Three recurring error types were observed: (i) entity omissions or substitutions, where key actors (e.g., organizations or locations) are missing or replaced by generic phrases; (ii) hallucinated details, where numbers, dates, or causal explanations not supported by the source article are introduced; and (iii) over-compression, where summaries become too short and drop essential background information. The second-phase optimization, which combines entity-aware ranking with mBART refinement, substantially reduces the first two categories (especially for named entities such as banks and cities) and slightly increases summary length, thereby mitigating over-compression while preserving the main event described in the article.

The experimental findings clearly demonstrate

Type	Text
Source Text	"বাংলাদেশ ব্যাংক ঘোষণা করেছে যে নতুন অর্থবছরের কৃষি খাতে ঋণের পরিমাণ বাড়ানো হবে। এর ফলে কৃষকরা স্বল্পসুদে ঋণ নিতে পারবেন।" (Bangladesh Bank has announced that the amount of loans in the agricultural sector will be increased in the new fiscal year. As a result, farmers will be able to take loans at low interest rates.)
Reference Summary (Human)	"কৃষি খাতে ঋণ বাড়াবে বাংলাদেশ ব্যাংক।" (Bangladesh Bank will increase loans in the agricultural sector.)
First-Level Summary	"বাংলাদেশ ব্যাংক নতুন অর্থবছরে কৃষি খাতে ঋণ বাড়াবে।" (Bangladesh Bank will increase agricultural loans in the new fiscal year.)
Optimized Summary	"বাংলাদেশ ব্যাংক কৃষকদের জন্য কৃষি খাতে ঋণ বাড়ানোর ঘোষণা দিয়েছে।" (Bangladesh Bank has announced an increase in agricultural loans for farmers.)

Table 3: Example of text sample comparison across phases

the effectiveness of the proposed two-phase optimization framework. Starting from the baseline model, which produced modest results, the incorporation of entity-aware embeddings and heuristic ranking in the first-level summarizer delivered measurable improvements in informativeness and coherence. The final optimization stage, leveraging NER reinforcement, redundancy filtering, and mBART rewriting, achieved the strongest performance across both automated metrics and human evaluation, substantially surpassing the baseline and intermediate systems. Automated results confirm notable gains in ROUGE, BLEU, and BERTScore, while human assessments highlight significant improvements in fluency, adequacy, coherence, and readability. These outcomes validate the central claim of this study: that integrating classical preprocessing, semantic embeddings, and instruction-tuned LLM refinement produces Bangla summaries that are not only concise and fluent but also factually reliable and stylistically aligned with human-written references.

5 Conclusion

5.1 Our Findings

The framework was evaluated on the Bengali Abstractive News Summarization Dataset (BANSData). Performance was assessed using ROUGE-1, ROUGE-L, BLEU, and BERTScore, complemented by human evaluation from three expert native Bangla linguists who rated fluency, adequacy, coherence, and readability on a five-point Likert scale.

The baseline model achieved ROUGE-1 0.30, ROUGE-L 0.31, BLEU 0.30, and BERTScore 0.52, with mean human ratings around 3.8. Incorporating entity-aware embeddings and a ranking mechanism in the first-level summarizer improved all metrics, especially BLEU (from 0.30 to 0.42) and BERTScore (from 0.52 to 0.55); human ratings increased to about 4.0. The Phase 2 optimized system, enhanced through NER-based entity preservation and mBART-driven semantic refinement, achieved ROUGE-1 0.54, ROUGE-L 0.36, BLEU 0.50, and BERTScore 0.88. Mean human evaluation scores exceeded 4.4 across all dimensions, confirming the framework's effectiveness in improving both lexical overlap and semantic fidelity.

Taken together, the automatic metrics and human evaluations confirm that the proposed two-phase framework substantially outperforms both the baseline and intermediate systems. Improvements averaged around +0.20 for ROUGE-1 and ROUGE-L, +0.20 for BLEU, and +0.36 for BERTScore. Human evaluation scores increased by approximately 0.6 points on the five-point Likert scale, reflecting consistent gains in fluency, adequacy, coherence, and readability. These findings validate the effectiveness of combining entity-aware draft generation with LLM-based optimization for producing high-quality Bangla abstractive summaries that achieve stronger lexical overlap and semantic fidelity.

5.2 Future Prospects

Further fine-tuning of mBART on domain-specific Bangla corpora could help mitigate trade-offs between brevity and semantic richness. Incorporating real-time data sources such as `bangla.bdnews24.com` may enhance adaptability to evolving news content. Expanding collaboration with a broader panel of linguists could also refine evaluation criteria and improve the balance between fluency and adequacy.

Limitations. The proposed framework is trained and evaluated on BANSData news articles, which may limit generalization to conversational or literary Bangla without domain adaptation. Nevertheless, the two-phase architecture itself is model- and domain-agnostic and can be extended to other Bangla genres (e.g., healthcare, education, social media) given appropriate training data and domain adaptation. Although entity preservation mitigates factual drift, no external fact-checking module is applied, and errors in Named Entity Recognition could propagate. Human evaluation is conducted with a small group of expert linguists using Likert-scale ratings, which, despite reporting inter-annotator agreement, remains inherently subjective. Furthermore, optimization with large language models increases computational cost, and efficiency-oriented techniques such as quantization or distillation are not yet explored.

References

Ajwad Abrar, Farzana Tabassum, and Sabbir Ahmed. 2024. Performance evaluation of large language models in bangla consumer health query summarization. In *Proceedings of the 2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 2748–2753, Dhaka, Bangladesh. IEEE.

Fahmida Afroja Hoque Barsha and Mohammed Nazim Uddin. 2023. Comparative analysis of BanglaT5 and pointer generator network for bengali abstractive story summarization. In *Proceedings of the 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, pages 84–88, Dhaka, Bangladesh. IEEE.

Prithwiraj Bhattacharjee, Avi Mallick, Md. Saiful Islam, and Marium-E-Jannat. 2021. Bengali abstractive news summarization (bans): A neural attention approach. In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*, pages 41–51, Singapore. Springer Singapore.

Washik Wali Faieaz, Sayma Jannat, Pronoy Kumar Mondal, Shahriar Shadman 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)*, pages 1–7, Chittagong, Bangladesh. IEEE.

Towhid Ahmed Foysal, Mohaimen Abid Mahadi, Md. Mahadi Hasan Nahid, and Ayesha Tasnim. 2021. Bangla-extrasum: Comparative analysis of different methods in automated extractive bengali text summarization. In *Proceedings of the 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pages 1–6, Dhaka, Bangladesh. IEEE.

Md. Majharul Haque, Suraiya Pervin, Anowar Hossain, and Zerina Begum. 2020. Approaches and trends of automatic bangla text summarization: Challenges and opportunities. *International Journal of Technology Diffusion (IJTD)*, 11(4):17–29.

Md. Nahid Hasan, Rafsan Bari Shafin, Marwa Khanom Nurtaj, Zeshan Ahmed, M. Saddam Hossain Khan, Rashedul Amin Tuhin, and Md. Mohsin Uddin. 2023. Implementation of bangla extractive update summarization task on BUSUM-BNLP dataset: A multi-document update summarization corpus. In *Proceedings of the 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–6, Istanbul, Turkiye. IEEE.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Samin Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Khan Md Hasib, Md. Atiqur Rahman, Mustavi Ibne Masum, Friso De Boer, Sami Azam, and Asif Karim. 2023. Bengali news abstractive summarization: T5 transformer and hybrid approach. In *Proceedings of the 2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 539–545, Port Macquarie, NSW, Australia. IEEE.

S. M. Afif Ibne Hayat, Avishek Das, and Mohammed Moshiul Hoque. 2023. Abstractive bengali text summarization using transformer-based learning. In *Proceedings of the 2023 6th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–6, Khulna, Bangladesh. IEEE.

Nahid Hossain, Mehedi Hasan Bijoy, Salekul Islam, and Swakkhar Shatabda. 2024. Panini: A transformer-based grammatical error correction method for bangla. *Neural Computing and Applications*, 36:3463–3477.

Nahid Hossain, Salekul Islam, and Mohammad Nurul Huda. 2021. Development of bangla spell and grammar checkers: Resource creation and evaluation. *IEEE Access*, 9:141079–141097.

Alam Khan, Sanjida Akter Ishita, Fariha Zaman, Ashiqul Islam Ashik, and Md. Moinul Hoque. 2023. Intelligent combination of approaches towards improved bangla text summarization. In *Proceedings of the 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, pages 1–6, Rajshahi, Bangladesh. IEEE.

Panagiotis Kouris, Georgios Alexandridis, and Andreas Stafylopatis. 2019. Abstractive text summarization based on deep learning and semantic content generalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5082–5092, Florence, Italy. Association for Computational Linguistics.

Namrata Kumari, Nikhil Sharma, and Pradeep Singh. 2023. Performance of optimizers in text summarization for news articles. In *Procedia Computer Science*, volume 218, pages 2430–2437. Elsevier.

Asif Ahammad Miazee, Tonmoy Roy, Md Robiul Islam, and Yeamin Safat. 2025. Abstractive text summarization for bangla language using NLP and machine learning approaches. *arXiv preprint arXiv:2501.15051*.

Aroni Saha Prapty, Md. Rifat Anwar, and K. M. Azharul Hasan. 2021. A rule-based parsing for bangla grammar pattern detection. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 319–331, Singapore. Springer Singapore.

Mizanur Rahman, Sajib Debnath, Masud Rana, Saydul Akbar Murad, Abu Jafar Md. Muzahid, Syed Zahidur Rashid, and Abdul Gafur. 2024. Bangla text summarization analysis using machine learning: An extractive approach. In *Proceedings of the 2nd Human Engineering Symposium (HUMENS 2023)*, pages 65–80, Singapore. Springer.

Faria Sultana, Md. Tahmid Hasan Fuad, Rahat Rizvi Rahman, Md. Hossain, Md. Ashraful Amin, Asif Rahman, and Amin Ahsan. 2024. How good are LM and LLMs in bangla newspaper article summarization? In *Pattern Recognition. ICPR 2024. Lecture Notes in Computer Science*, volume 15320, pages 77–91, Cham. Springer Nature.

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)*, pages 35–45, Singapore. Association for Computational Linguistics.

Ha Nguyen Thi Thu. 2014. An optimization text summarization method based on naïve bayes and topic word for single syllable language. *Applied Mathematical Sciences*, 8(3):99–115.

Muhammad Hafizul H. Wahab, Nor Hafiza Ali, Nor Asilah Wati Abdul Hamid, Shamala K. Subramaniam, Rohaya Latip, and Mohamed Othman. 2024. A review on optimization-based automatic text summarization approach. *IEEE Access*, 12:4892–4909.