Distinguishing Repetition Disfluency from Morphological Reduplication in Bangla ASR Transcripts: A Novel Corpus and Benchmarking Analysis

Zaara Zabeen Arpa [1] Sadnam Sakib Apurbo [1] Nazia Karim Khan Oishee [1] Ajwad Abrar [1]

Abstract

Automatic Speech Recognition (ASR) transcripts, especially in low-resource languages like Bangla, contain a critical ambiguity: word-word repetitions can be either Repetition Disfluency (unintentional ASR error/hesitation) or Morphological Reduplication (a deliberate grammatical construct). Standard disfluency correction fails by erroneously deleting valid linguistic information. To address this, we introduce the first publicly available Bangla corpus of 20,000 sentences, manually annotated to explicitly distinguish between these two phenomena in noisy ASR transcripts. We benchmark this novel resource using two paradigms: state-of-the-art multilingual Large Language Models (LLMs) and task-specific fine-tuning of encoder models. LLMs achieve competitive performance (up to 82.68% accuracy) with few-shot prompting. However, fine-tuning proves superior, with the language-specific BanglaBERT model achieving the highest accuracy of 84.78% and an F1 score of 0.677. This establishes a strong, linguistically-informed baseline and provides essential data for developing sophisticated, semantic-preserving text normalization systems for Bangla.

1. Introduction

1.1. ASR Pitfalls: Disfluency and Repetition

Automatic Speech Recognition (ASR) is now integral to digital interaction, driving applications from virtual assistants to automated subtitles on platforms like YouTube (Dykes et al., 2023). Despite its wide adoption and significant advancements, ASR performance remains imperfect, particularly in Large Vocabulary Continuous Speech Recognition (LVCSR) and under real-world conditions like background noise or diverse accents (Errattahi et al., 2018; Romana et al., 2024).

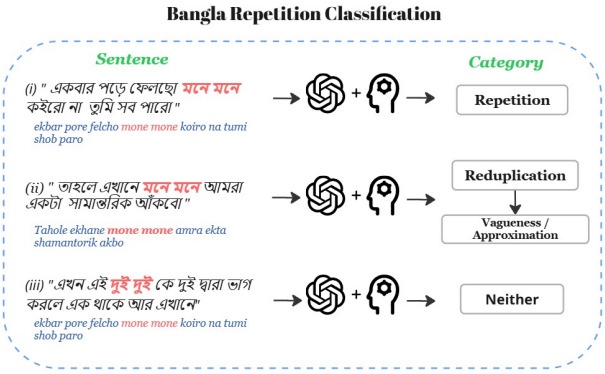

Figure 1. Illustration of the Bangla Repetition Classification task, highlighting the distinction between unintentional disfluencies (Repetition), grammatical forms (Reduplication), and coincidental occurrences (Neither).

This often results in a significant Word Error Rate (WER). A major, persistent source of error stems from speech disfluencies, which are interruptions in the smooth flow of speech, including filled pauses, hesitations, self-corrections, and most relevantly, repetitions (Romana et al., 2024). Disfluencies are natural and frequent; one study found a 50% probability of a disfluency in a 10-13 word sentence (Jamshid Lou & Johnson, 2020b). Their presence creates noisy transcripts that are difficult to read and detrimental to downstream Natural Language Processing (NLP) tasks such as machine translation or information extraction (Romana et al., 2024; Jamshid Lou & Johnson, 2020a). Consequently, automatic Disfluency Correction (DC) is a critical research area, aiming to "clean" ASR outputs (Errattahi et al., 2018). High-quality DC corpora, such as DISCO, have enabled benchmark F1 scores up to 94.29% in languages

[1]Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. Correspondence to: Ajwad Abrar <ajwadabrar@iut-dhaka.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.
Copyright 2026 by the author(s).

like Hindi, confirming DC's vital post-processing role (Bhat et al., 2023b).

1.2. Bangla Ambiguity: Disfluency vs. Reduplication

The conventional DC approach treats repetitions as universal "noise" to be removed. This fails in languages with specific morphological properties that structurally align with disfluent phenomena. This paper addresses a critical ambiguity in Bangla, an Indo-Aryan language with over 270 million speakers (Ridoy et al., 2025). In Bangla ASR transcripts, the identical surface form *word-word* can represent two phenomena with opposing grammatical implications:

1. Repetition Disfluency: An unintentional, non-grammatical repetition (speaker hesitation or ASR error), often described by the Reparandum-Interregnum-Repair (RiR) framework (Bhat et al., 2023b; Ahmad et al., 2025). Example: "এই যে পাঁচ তারিখ থেকে শুরু করে করে তোমার 12 তারিখ পর্যন্ত।" (...shuru kore kore...). Here, the repeated "করে" (kore) is an error and should be deleted.

2. Morphological Reduplication: A deliberate, rule-governed, and grammatically significant process where a word is repeated to convey a specific semantic nuance, such as continuity, iterativity, intensity, or plurality (Rácová, 2013; Abbi, 1992). Example: "অংকগুলো করে করে আমরা একটু আন্ডারস্ট্যান্ডিং ডেভেলপ করা চেষ্টা করবো" (Onkogulo kore kore...). The repeated phrase "করে করে" (kore kore) is a crucial construction conveying an iterative nature and must be preserved.

This structural ambiguity, visually detailed in Figure 1, presents a formidable challenge. Generic disfluency detection models designed with a "subtractive" philosophy would fail by erroneously stripping away valid linguistic information, catastrophically altering the semantic content of the text. Resolving this requires a fine-grained, context-aware classification model.

1.3. The Low-Resource Data Gap

Developing models to resolve such language-specific ambiguities requires large-scale, high-quality annotated data. Despite its massive speaker base, Bangla is a low-resource language in NLP (Ridoy et al., 2025) due to a scarcity of standardized, publicly available datasets. For the specific task of distinguishing repetition disfluency from morphological reduplication in Bangla, no publicly available annotated corpus existed prior to this work. This resource gap has been the primary impediment to developing and rigorously evaluating computational systems for this task, hindering the shift from generic, one-size-fits-all NLP solutions toward models sensitive to the unique grammatical structures of low-resource languages.

1.4. Contributions

This paper addresses this critical resource and research gap by providing the necessary data and establishing strong performance baselines. The primary contributions are threefold:

1. Corpus Creation: We introduce the first publicly available,[1] Bangla corpus of 20,000 sentences, manually annotated to explicitly distinguish between Repetition Disfluency and Morphological Reduplication in noisy ASR transcripts. Furthermore, we provide a fine-grained linguistic analysis by subcategorizing all Morphological Reduplication instances into nine distinct semantic and functional classes.

2. LLM Benchmarking: We benchmark state-of-the-art multilingual Large Language Models (LLMs) (GPT, Gemini, Claude families) under zero-shot, one-shot, and few-shot prompting. LLMs achieve a competitive performance up to 82.68% accuracy with few-shot prompting.

3. Fine-Tuning Analysis: We empirically demonstrate the superiority of task-specific fine-tuning. The language-specific BanglaBERT (Bhattacharjee et al., 2022) model achieves the highest performance with an accuracy of 84.78% and an F1 score of 0.677, establishing a strong, linguistically-informed baseline for developing semantic-preserving text normalization systems for Bangla.

2. Related Works

Our research is situated at the intersection of speech processing, computational linguistics, and low-resource NLP. We contextualize our contribution by reviewing the distinct treatment of repetition as an error in disfluency correction versus a meaningful construct in morphological reduplication, and by outlining the standard methodological paradigms our work builds upon.

[1] https://huggingface.co/datasets/ajwad-abrar/BanglaReDup

2.1. The Dichotomy of Repetition: Disfluency vs. Reduplication

Computational Disfluency Correction (DC) is a critical post-processing step for ASR, designed to improve transcript readability by identifying and removing phenomena like filled pauses, self-corrections, and repetitions (Romana et al., 2024). The field has evolved from classic sequence tagging to sophisticated Transformer-based architectures, with large-scale corpora like DISCO enabling high F1 scores in high-resource languages (Bhat et al., 2023b; Jamshid Lou & Johnson, 2020b). For low-resource languages, including Bengali, the lack of labeled data has spurred techniques like zero-shot learning with multilingual encoders and synthetic data augmentation via adversarial training to improve performance (Kundu et al., 2022; Bhat et al., 2023a; Wang et al., 2022).

However, a fundamental limitation of the dominant DC paradigm is its inherently subtractive nature that treats all repetitions as "noise" to be deleted (Jamshid Lou & Johnson, 2020b;a). This approach is incompatible with languages like Bangla, where repetition is also a productive grammatical device. In linguistics, morphological reduplication is a rule-governed process where a word is repeated to encode specific semantic nuances, such as continuity, iterativity, or intensity (Rácová, 2013; Abbi, 1992). This creates a critical structural ambiguity where the surface form 'word-word' can be either an error or a meaningful linguistic construct.

This specific challenge is recognized across the Indo-Aryan language family. Recent parallel work has successfully benchmarked this classification task in Hindi, Marathi, and Telugu, achieving Macro F1 scores up to 85.62% and demonstrating the necessity of context-aware models that can distinguish grammatical reduplication from disfluent structures like the Reparandum-Interregnum-Repair (RiR) pattern (Ahmad et al., 2025). While early computational work on reduplication in Indic languages relied on rule-based systems or finite-state transducers (Chakraborty & Bandyopadhyay, 2010; Dolatian & Heinz, 2019), our work addresses this problem using modern neural architectures. We bridge the gap between the subtractive ASR-processing paradigm and principled linguistic analysis by creating a resource to train models for this nuanced classification task.

2.2. Methodological Context and Modeling Paradigms

Developing robust NLP solutions for Bangla is hindered by a scarcity of standardized datasets, a common challenge for low-resource languages despite their large speaker populations (Ridoy et al., 2025). This has motivated broad efforts to create foundational resources and models for the Indic language family (Kakwani et al., 2020). Our approach to corpus creation aligns with pragmatic solutions to this data gap: we leverage noisy, auto-generated ASR transcripts from YouTube (Dykes et al., 2023). This strategy is effective because the inherent flaws of ASR systems provide a naturalistic distribution of the very phenomena knowingly spurious repetitions, speaker hesitations, and correctly transcribed reduplications, required to train a robust real-world classifier.

For the classification task itself, we evaluate the two dominant paradigms for applying pre-trained models: in-context learning (Zhou et al., 2024) via prompting and task-specific fine-tuning (Liu et al., 2023). The former tests the ability of massive LLMs to perform the task with zero or few examples, while the latter adapts the weights of smaller, pre-trained encoder models to the specific dataset. Our experiments provide a direct comparison of these approaches, utilizing both general multilingual models (mBERT, XLM-RoBERTa) and the language-specific BanglaBERT (Bhattacharjee et al., 2022) to establish a strong, linguistically-informed baseline.

3. Methodology

Our methodological framework is structured around three key phases: Corpus Creation, LLM Benchmarking (Prompting), and Task-Specific Fine-Tuning. The comprehensive workflow for the corpus creation phase is illustrated in Figure 2. This overall design establishes a strong performance baseline by rigorously comparing the capabilities of in-context learning against transfer learning for this fine-grained linguistic classification task.

3.1. Corpus Creation: The Bangla Repetition Corpus

The Bangla Repetition Corpus was synthesized from real-world, noisy Automatic Speech Recognition (ASR) transcripts, ensuring a naturalistic distribution of both errors and grammatical forms. This process involved four steps: Scalable Data Acquisition, Automated Filtering, Expert Annotation, and Fine-Grained Sub-categorization.

3.1.1. Scalable Data Acquisition

We selected the top 10 popular Bangla educational YouTube channels as the data source to ensure a high volume of continuous, spontaneous spoken Bengali.

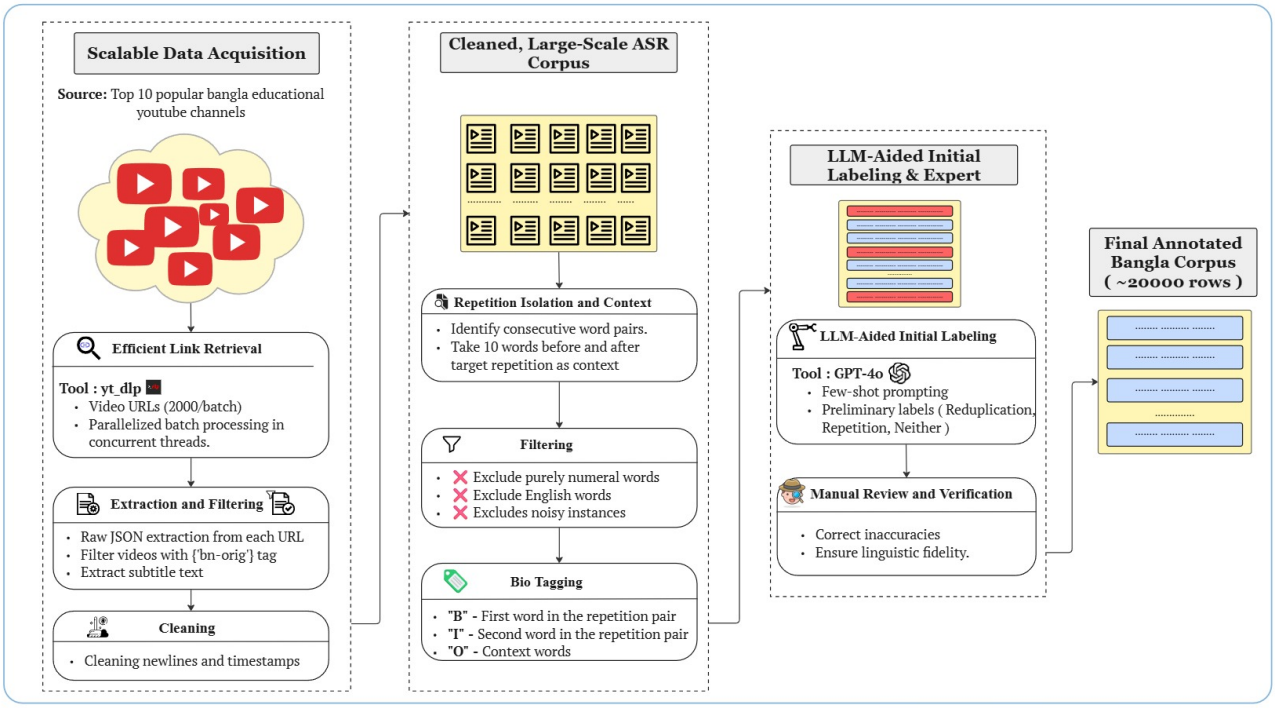

Figure 2. The end-to-end pipeline for creating the Bangla Repetition Corpus. The workflow begins with scalable data acquisition from YouTube ASR transcripts, followed by automated filtering and context extraction. The core annotation phase follows a hybrid approach, combining an LLM for initial labeling with expert linguists for final verification, resulting in a gold-standard corpus of 20,000 sentences.

1. Efficient Link Retrieval: Video Uniform Resource Locators (URLs) were collected using the yt-dlp utility. To maximize throughput, the retrieval process was parallelized using concurrent threads. We processed content from the channels' "videos" tabs in batches of 2000 to quickly compile a comprehensive list of video links.

2. ASR Transcript Filtering: Only videos confirmed to contain Bangla ASR transcripts were retained. This was programmatically verified by checking for the presence of the bn-orig tag in the subtitle manifest, confirming the noisy nature of the source data.

3. Subtitle Extraction and Cleaning: Subtitles were downloaded and extracted from the raw JSON format using a parallel process. The text segments were joined, and noise reduction included replacing newlines and standardizing whitespace.

3.1.2. Automated Filtering and Structural Tagging

From the cleaned, large-scale corpus of ASR text, we automatically extracted the ambiguous cases of contiguous word repetition:

1. Repetition Isolation: The corpus was processed in memory-efficient chunks. Consecutive, identical word pairs ($\text{word}_i = \text{word}_{i-1}$) were identified after tokenization using a regular expression that explicitly handles Bengali, English, and numerical tokens ([0980-09FF]+|[a-zA-Z0-9]+), while excluding purely numerical repetitions.

2. Context Window Formulation: Each repetition instance was extracted alongside a fixed, symmetric context window of 10 preceding words and 10 following words. This step resulted in a set of approximately $29,000$ sentences containing candidate repetition instances.

3. Initial BIO Tagging: To facilitate manual annotation and set up the task for supervised models, the repeated words within the context window were pre-tagged using a BIO scheme: the first occurrence was marked B (Beginning) and the second was marked I (Inside/Continuation).

3.1.3. LLM-Aided Initial Labeling and Expert Annotation

The large set of approximately 29,000 filtered sentences underwent a two-stage labeling process, combining the scalability of generative models with the precision of expert human review.

1. LLM-Aided Initial Labeling: To accelerate the annotation of the vast corpus, the sentences were first processed using the leading commercial model, GPT-4o, as an initial categorization engine. The model utilized the structured, few-shot prompt strategy, providing preliminary labels (Reduplication, Repetition, or Neither) for the entire set.
2. Expert Manual Review and Verification: The corpus with the preliminary LLM labels was then subjected to a rigorous manual review by expert Bengali speakers. This critical step served to correct any inaccuracies introduced by the LLM and to ensure linguistic fidelity, particularly for nuanced cases where the distinction between a complex reduplication form and a speaker hesitation was ambiguous.
3. Final Corpus Selection: Following the comprehensive manual verification and cleaning, all sentences that could not be unambiguously classified (often due to extreme ASR noise or context fragmentation) were discarded. This process finalized the core corpus, resulting in 20,000 gold-standard sentences used for training and evaluation.

The final annotated corpus of 20,000 sentences was categorized into three mutually exclusive labels:

1. Reduplication (Grammatical): The repetition is an intentional, rule-governed morphological process that conveys semantic nuances such as iterativity, continuity, intensity, or plurality. Example: "অংকগুলো করে করে আমরা একটু আন্ডারস্ট্যান্ডিং ডেভেলপ করা চেষ্টা করবো" (Onkogulo kore kore amra...)
2. Repetition (Disfluency): The repetition is an unintentional error, either a speaker-induced hesitation or an ASR transcription error. Example: "এই যে পাঁচ তারিখ থেকে শুরু করে করে তোমার 12 তারিখ পর্যন্ত।" (...shuru kore kore tomar...)
3. Neither: Coincidental repetitions that are not clear disfluencies or productive reduplications.

The final distribution of the annotated corpus, which shows a significant imbalance, is presented in Table 1.

Category	Percentage
Reduplication	66.3%
Repetition	32.9%
Neither	0.8%

Table 1. Category Distribution of the Annotated Bangla Corpus

3.1.4. Fine-Grained Sub-categorization of Reduplication

Following the primary classification, all instances identified as Morphological Reduplication underwent a second stage of fine-grained annotation to determine their specific semantic function. This was accomplished using a Large Language Model guided by a carefully constructed few-shot prompt that defined nine distinct subcategories: Intensity/Emphasis, Frequency/Iteration, Continuity/Ongoing Action, Plurality/Multiplicity, Distributive/Separateness, Vagueness/Approximation, Echo Word/Rhyming, Reciprocal/Correlative, and Onomatopoeia. For instance, in the sentence "...এক্স এর ভ্যালু পেয়ে গেছি কত কত বলো..." (...we got the values of x, tell me what what...), the repeated word "কত কত" (koto koto) implies an iterative query for multiple values, leading to its classification as Frequency / Iteration. This two-tiered annotation process enriches the corpus, providing a detailed linguistic layer for future research.

3.2. Experimental Setup and Baselines

The evaluation was conducted across two distinct experimental setups on the held-out test set of 335 sentences, comparing the efficacy of in-context learning against transfer learning.

3.2.1. LLM Benchmarking with Prompting Strategies

We established a prompting baseline by evaluating seven state-of-the-art Large Language Models (LLMs) on this classification task. This set included leading proprietary models (GPT-4o, Claude 4, and Gemini 2.5 Flash) and several prominent open-source alternatives (Gemma 3, Mistral 7b instruct, Llama 3 8b Instruct, and Phi-4). The models were tested under Zero-shot, One-shot, and Few-shot ($N \leq 5$) conditions.

Our prompting strategy utilized a strictly controlled setup:

- Inference Control: The temperature was set to 0.1 to favor deterministic and stable classification outputs.

Models	Zero-shot (%)	One-shot (%)	Few-shot (%)
Claude 4 sonnet	76.41	80.29	82.68
GPT-4o	78.50	80.50	82.10
Gemini 2.5 Flash	78.51	76.72	81.49
Phi-4	61.19	60.30	62.39
Gemma 3.4b	63.88	61.10	46.20
Llama 3 8b Instruct	56.12	54.33	53.73
Mistral 7b instruct	43.88	66.27	62.09

Table 2. Accuracy (%) of Different Prompting Techniques on the Bangla Corpus. Bold values indicate the peak accuracy achieved by each model across the different prompting strategies.

- Structured Prediction: All prompts enforced a Structured JSON-in, JSON-out format, requiring the model to output a single, valid JSON object containing only the predicted category, which minimizes parsing errors and enforces a consistent response structure.

- Explicit Context: For few-shot tests, the prompt included explicit linguistic definitions and examples for the three target categories (Reduplication, Repetition, and Neither) to guide the LLMs' in-context learning capability.

3.2.2. Task-Specific Fine-Tuning of Encoder Models

We established robust performance baselines by conducting task-specific fine-tuning on three prominent Transformer-based encoder models: a Bangla-specific model and two high-performing multilingual models. The task was framed as a three-way sequence classification (sentence-level classification into Reduplication, Repetition, or Neither).

- Models: We selected BanglaBERT (a language-specific model pre-trained on a vast Bengali corpus), XLM-RoBERTa (base), and mBERT (two widely-used multilingual models).

- Training Parameters: All models were fine-tuned for 3 epochs with a Batch Size of 16 and a Learning Rate of 2e-5. A Weight Decay of 0.01 was applied, and the Max Length was set to 128 tokens.

This approach directly compares the efficacy of using general multilingual pre-trained knowledge (mBERT/XLM-R) against specialized language pre-training (BanglaBERT) when adapting to a novel, linguistically-sensitive classification task on limited, noisy data.

4. Results

4.1. LLM Benchmarking with Prompting Strategies

Table 2 presents the accuracy of different multilingual LLMs across the three prompting strategies.

4.1.1. Prompting Strategy Effectiveness

The zero-shot performance of the leading LLMs (e.g., Gemini 2.5 Flash at 78.51%) demonstrates that massive multilingual pre-training imparts a strong baseline capability to resolve word repetition ambiguity, likely leveraging latent knowledge across the Indo-European family. For the top models, few-shot prompting was the most effective method, consistently boosting accuracy to over 81%. This confirms that explicit in-context examples are necessary to guide the LLMs toward the subtle grammatical cues that distinguish morphological reduplication from disfluency. Claude 4, for instance, achieved the highest LLM performance at 82.68%. However, providing examples was inconsistent for the lower-tier models, sometimes degrading performance, which suggests that their internal representations are less robustly aligned with the task, and their ability to generalize from few-shot examples is limited.

4.1.2. Model Ranking

The consistently high performance of Claude 4, GPT-4o, and Gemini 2.5 Flash across all prompting configurations established them as the top-performing models (Table 2). Conversely, the relatively low and inconsistent results from open-source models like Gemma 3.4b and Llama 3 8b Instruct underscored the challenge of generalizing complex linguistic rules in a low-resource setting without substantial specialized pre-training.

- Top-Performing: Claude 4, GPT-4o, Gemini 2.5 Flash.

- Mid-Tier: Mistral 7b instruct, Phi-4.

Model Name	Type	Accuracy (%)	Precision	Recall	F1 Score
BanglaBERT	Base	59.70	0.394	0.395	0.381
	Fine Tuned	84.78	0.901	0.646	0.677
XLM-RoBERTa	Base	37.31	0.267	0.331	0.190
	Fine Tuned	83.28	0.556	0.580	0.566
mBert	Base	37.31	0.124	0.333	0.181
	Fine Tuned	83.28	0.553	0.581	0.565

Table 3. Fine-tuning Results: Accuracy, Precision, Recall, and F1 Score. The fine-tuning process yields substantial gains in accuracy for all models (approx. 24-46 percentage points). BanglaBERT emerges as the strongest performer, achieving the highest Accuracy (84.78%) and a superior F1 Score (0.677). The high Precision (0.901) achieved by BanglaBERT is crucial, indicating a strong ability to preserve grammatically meaningful Reduplication instances, prioritizing semantic integrity over comprehensive disfluency removal.

- Low-Performing: Gemma 3 4b, Llama 3 8b Instruct.

4.2. Impact of Task-Specific Fine-Tuning

Fine-tuning the encoder models on our custom Bangla dataset led to substantial improvements in all metrics, significantly surpassing the highest performance achieved by the LLMs through prompting. Figure 3 illustrates the sharp gain in accuracy for all three models after fine-tuning.

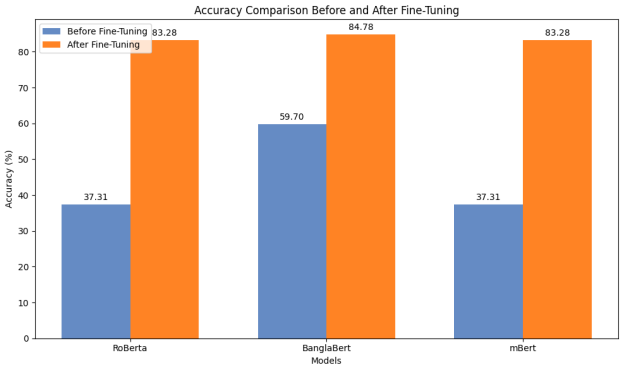

Figure 3. Accuracy Comparison Before and After Fine-Tuning

The numerical results for accuracy and other key metrics are detailed in Table 3.

4.2.1. Key Fine-Tuning Findings

- BanglaBERT (Bhattacharjee et al., 2022) Superiority and Cross-Linguistic Context: BanglaBERT achieved the highest performance across the board after fine-tuning, with an accuracy of 84.78% and the highest precision (0.901) and F1 score (0.677). This result is competitive and consistent with state-of-the-art Macro F1 scores achieved in parallel research on this specific reduplication/repetition classification task in related Indo-Aryan languages, such as Hindi (up to 85.62%) and Marathi (up to 84.82%) (Ahmad et al., 2025). This highlights the value of using a language-specific model for a nuanced task in a low-resource language.

- Fine-Tuning vs. Prompting: The best fine-tuned model (BanglaBERT, 84.78% accuracy) significantly outperformed the best-prompted LLM (Claude 4, 82.68% accuracy), demonstrating that for this specific, linguistically-motivated classification task, the comprehensive parameter updates of fine-tuning are more effective than in-context learning.

- Metric Disparity Analysis: A key observation is the substantial disparity between Precision (0.901) and Recall (0.646) for the Fine-Tuned BanglaBERT model. This disparity reflects the consequence of the highly imbalanced dataset (66.3% Reduplication vs. 32.9% Repetition). The high precision is highly desirable for normalization, as it indicates the model is extremely conservative, successfully avoiding False Positives (i.e., erroneously deleting meaningful Reduplication instances). Conversely, the lower recall shows that the model still misses a significant number of true Repetition Disfluency instances (False Negatives), allowing noise to remain in the transcript. This conservatism represents a strategic trade-off, prioritizing semantic preservation over comprehensive noise removal.

- Consistent Gains: XLM-RoBERTa and mBERT both showed similar and substantial gains, reaching an accuracy of 83.28%. However, their F1 scores remained notably lower than BanglaBERT, reinforcing the advantage of specialized language pre-training.

5. Conclusion

This paper addresses the critical ambiguity between grammatical Morphological Reduplication and erroneous Repetition Disfluency in Bangla ASR transcripts. To solve this, we introduce the first publicly available, annotated corpus for this classification task. Our experiments demonstrate that task-specific fine-tuning is superior to few-shot prompting of large language models. The language-specific BanglaBERT model established the strongest performance baseline, achieving an accuracy of 84.78%. This work provides the essential data and a validated benchmark, paving the way for developing robust, semantic-preserving text normalization systems for Bangla.

References

Abbi, A. Reduplication in south asian languages : an areal, typological, and historical study. 1992. URL https://api.semanticscholar.org/CorpusID:127735640.

Ahmad, A. A., Mothika, K. G., and Bhattacharyya, P. Looks can be deceptive: Distinguishing repetition disfluency from reduplication. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S. (eds.), Proceedings of the 31st International Conference on Computational Linguistics, pp. 214–229, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.15/.

Bhat, V., Jyothi, P., and Bhattacharyya, P. Adversarial training for low-resource disfluency correction. In Findings of the Association for Computational Linguistics: ACL 2023, pp. 8112–8122, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.514.

Bhat, V., Jyothi, P., and Bhattacharyya, P. DISCO: A large scale human annotated corpus for disfluency correction in indo-european languages. In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 12833–12857, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.855.

Bhattacharjee, A., Hasan, T., Ahmad, W., Mubasshir, K. S., Islam, M. S., Iqbal, A., Rahman, M. S., and Shahriyar, R. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), Findings of the Association for Computational Linguistics: NAACL 2022, pp. 1318–1327, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.98. URL https://aclanthology.org/2022.findings-naacl.98/.

Chakraborty, T. and Bandyopadhyay, S. Identification of reduplication in bengali corpus and their semantic analysis: A rule based approach. In MWE@COLING, 2010. URL https://api.semanticscholar.org/CorpusID:12590683.

Dolatian, H. and Heinz, J. Computational models of morphological copying. 2019. URL https://iacs.stonybrook.edu/_pdf/hossep_dolatian_2019RD.pdf.

Dykes, N., Wilson, A., and Uhrig, P. A pipeline for the creation of multimodal corpora from YouTube videos. In Aggarwal, P., Ala{\c{c}}am, {\"O}., Silberer, C., Zarrie{\ss}, S., and Zesch, T. (eds.), Proceedings of the 1st Workshop on Linguistic Insights from and for Multimodal Language Processing, pp. 1–5, Ingolstadt, Germany, September 2023. Association for Computational Lingustics. URL https://aclanthology.org/2023.limo-1.1/.

Errattahi, R., El Hannani, A., and Ouahmane, H. Automatic speech recognition errors detection and correction: A review. Procedia Computer Science, 128:32–37, 2018. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2018.03.005. URL https://www.sciencedirect.com/science/article/pii/S1877050918302187. 1st International Conference on Natural Language and Speech Processing.

Jamshid Lou, P. and Johnson, M. End-to-end speech recognition and disfluency removal. In Cohn, T., He, Y., and Liu, Y. (eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2051–2061, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.186. URL https://aclanthology.org/2020.findings-emnlp.186/.

Jamshid Lou, P. and Johnson, M. Improving disfluency detection by self-training a self-attentive model. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3754–3763, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.346. URL https://aclanthology.org/2020.acl-main.346/.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. M., and Kumar, P. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Cohn, T., He, Y., and Liu, Y. (eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.445. URL https://aclanthology.org/2020.findings-emnlp.445/.

Kundu, R., Jyothi, P., and Bhattacharyya, P. Zero-shot disfluency detection for Indian languages. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H. (eds.), Proceedings of the 29th International Conference on Computational Linguistics, pp. 4442–4454, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.392/.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Comput. Surv., 55(9), January 2023. ISSN 0360-0300. doi: 10.1145/3560815. URL https://doi.org/10.1145/3560815.

Ridoy, M. S. I., Akter, S., and Rahman, M. A. Adaptability of asr models on low-resource language: A comparative study of whisper and wav2vec-bert on bangla. 2025.

Romana, A., Koishida, K., and Provost, E. M. Automatic disfluency detection from untranscribed speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32:4727–4740, 2024. doi: 10.1109/TASLP.2024.3485465.

Rácová, A. Reduplication of verbal forms in bengali. 22, 12 2013. URL https://www.researchgate.net/publication/281318711_Reduplication_of_verbal_forms_in_Bengali.

Wang, Z., Wang, Y., Wang, S., and Che, W. Adaptive unsupervised self-training for disfluency detection. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H. (eds.), Proceedings of the 29th International Conference on Computational Linguistics, pp. 7209–7218, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.632/.

Zhou, Y., Li, J., Xiang, Y., Yan, H., Gui, L., and He, Y. The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 14365–14378, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.795. URL https://aclanthology.org/2024.emnlp-main.795/.