Design and Evaluation of a Robust Watermarking Algorithm for Bangla Text Generation Using Large Language Models

A N M Zahid Hossain Milkan	**200041202**
Amit Bin Tariqul	**200041214**
Sahab Al Chowdhury	**200041255**

Supervised By

Dr. Kamrul Hasan
Professor

Dr. Hasan Mahmud
Professor

Syed Rifat Raiyan
Lecturer

Systems and Software Lab (SSL)
Dept. of Computer Science & Engineering, Islamic University of Technology

Introduction

LLM Watermarking

Definition:
LLM watermarking is the process of embedding hidden, machine-detectable patterns in AI-generated text to identify its origin while remaining invisible to humans.

Example:

Here is a prompt given to a LLM. It generates both non-watermarked as well as watermarked versions. Non watermarked version has much lower percentage of green (watermark candidate) tokens than the watermarked version. These green tokens contribute to distinguishing between human and AI generated text.

Goal:
Attempt to **emulate** watermarking capabilities in Bangla LLMs to ensure secure and traceable text generation.

Prompt

The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:

No Watermark

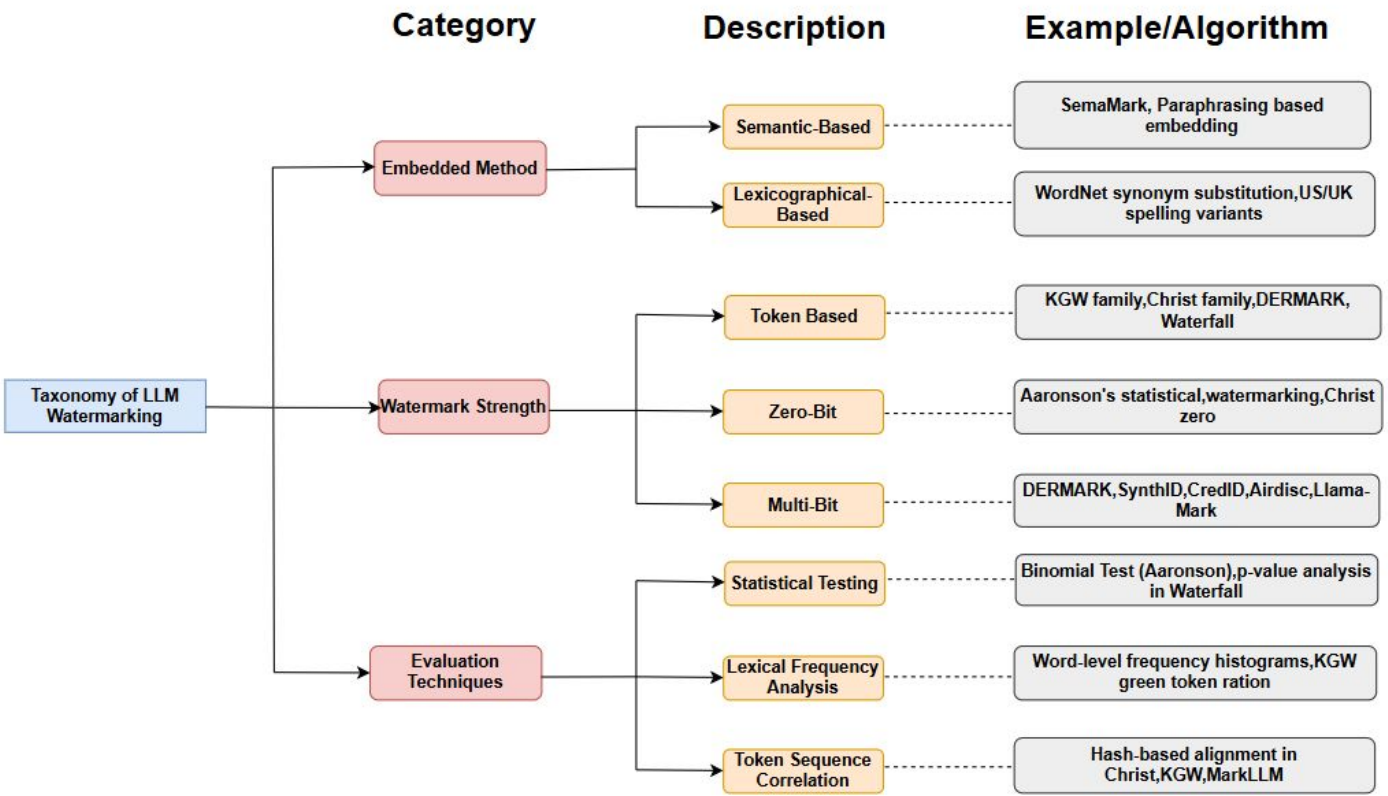
Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words)
Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet

Watermarked

-minimal marginal probability for a detection attempt.
-Good speech frequency and energy rate reduction.
-messages indiscernible to humans.
-easy for humans to verify.

A Taxonomy of Watermarking Algorithms

Watermarking algorithms can be classified into various types based on specific criteria.



Objectives

LLM Watermarking

Develop a watermark algorithm specifically adapted for ***Bangla LLM*** text generation

Our focus is to -

- Create a lightweight and reliable watermarking method for Bangla outputs.
- Investigate and categorize potential attacks on watermarks with corresponding defense strategies
- Conduct linguistic analysis of morphology and syntax for effective watermark integration
- Perform comparative analysis of existing watermarking techniques for **Bangla** applications
- Identify and analyze resource gaps and practical challenges in this field

Motivation

Our Inspiration for pursuing this topic

- Bengali, ranked as the **7th most spoken language** globally with over **300 million speakers**, exhibits complex morphology, compound verb constructions, and diverse linguistic features, warranting focused research attention.

- Although there exists approaches for robust LLM watermarking for different languages, **no attempts have been made** to implement the same for Bangla.

- LLM watermarking will prevent spread of misinformation by empowering fact-checking initiatives and sense of accountability.

- South Asia lacks legal frameworks for AI-generated content. This work could establish precedents for **protecting Bangla IP** (Intellectual Property) in industries like media, literature and tech.

- **Bangla-Medium** schools and universities **lack tools** to detect AI plagiarism. This could pioneer academic integrity solutions for the domain.

Problem Formulation

How watermarking is implemented & Detected

We have a Large Language Model (LLM) ***M*** capable of producing Bangla text when given a prompt ***P.*** Each model has a vocabulary ***V*** from which it generates the next token $\mathbf{t_i}$ by factoring in the probability based on previous tokens $\mathbf{t_{i-1}}$ to $\mathbf{t_0}$ (a certain length).

The goal of the watermarking process is to tweak the model to produce certain output tokens from either a fixed list of tokens or from a different distribution thus allowing watermark to be embedded in the overall output.This output when later used anywhere else say a text ***X*** , can be traced back to the original LLM ***M*** which produced it by detecting the pattern or presence of those watermarked tokens through a statistical test or **p-value** test.

Literature Review

Protecting Intellectual Property of Language Generation APIs with Lexical Watermark[8]

Xuanli He , Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, Chenguang Wang (Microsoft Research Asia,UC Berkeley) Jun 2021

This paper introduces a **lexical watermarking** method to protect commercial NLG APIs (e.g., machine translation, summarization) from **Model Extraction Attacks(MEA)**, where adversaries tries to copy machine learning model by repeatedly sending it inputs and using the outputs to train their own version,to avoid payment.

Contributions

- This paper is the **first to** adapt watermarking to generative text models
- **Watermarking via Lexical Substitution:** Words are replaced with rare synonyms or US/UK spelling variants using a secret hash function (e.g., "great" → "outstanding", "color" → "colour").

Advantages

- Only **10% of the output** needs to be watermarked for detection.
- **Preserves text quality** by ensuring minimal distortion

Disadvantages

- **Limited to lexical changes**: Struggles with low-entropy text (e.g., code, formulas).
- ❌ **Not designed for large LMs**: Focuses on task-specific models (translation/summarization).

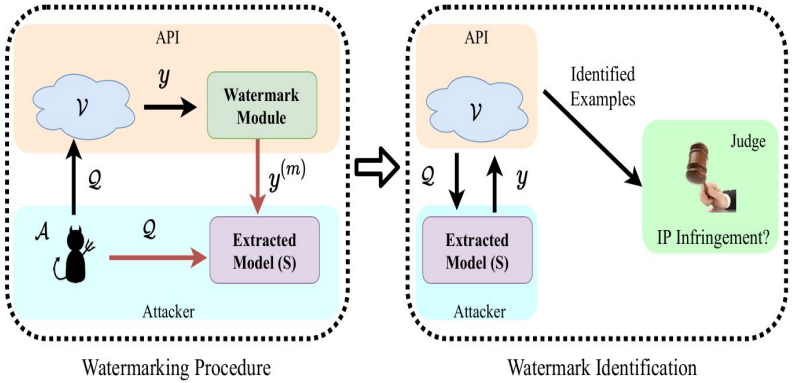


Figure: The left figure shows that the output y of queries are watermarked before answering end-users. At the $\mathcal{Q}$ watermark identification phase, the victim $\mathcal{V}$ first queries the suspicious model to obtain some text y.Then it will be examined by $\mathcal{V}$ and judged for the ownership claim.

Literature Review

A Watermark for Large Language Models[3][7]

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, Tom Goldstein in Proceedings of the 40th International Conference on Machine Learning, PMLR 202:17061-17084, 2023.

To solve the limitation of previous paper, this paper proposes a **token-level watermarking** method for LLMs by **biasing output** toward **"green-listed"** tokens, detectable via statistical tests while avoiding **"red listed"** tokens. While robust to minor edits, it's vulnerable to paraphrasing but balances quality and detectability with theoretical guarantees.

Contributions:

- First **practical** LLM watermark with **model free** detection.(only need hash and RNG)
- Proposed **soft vs. hard watermarking** to balance **quality and robustness**.

Advantages:

- No model retraining and public detection
- Quality preservation and great statistical guarantee

Drawbacks:

- **Vulnerable to low entropy** sequences and slow.
- **Vulnerable to paraphrasing** and token based attacks

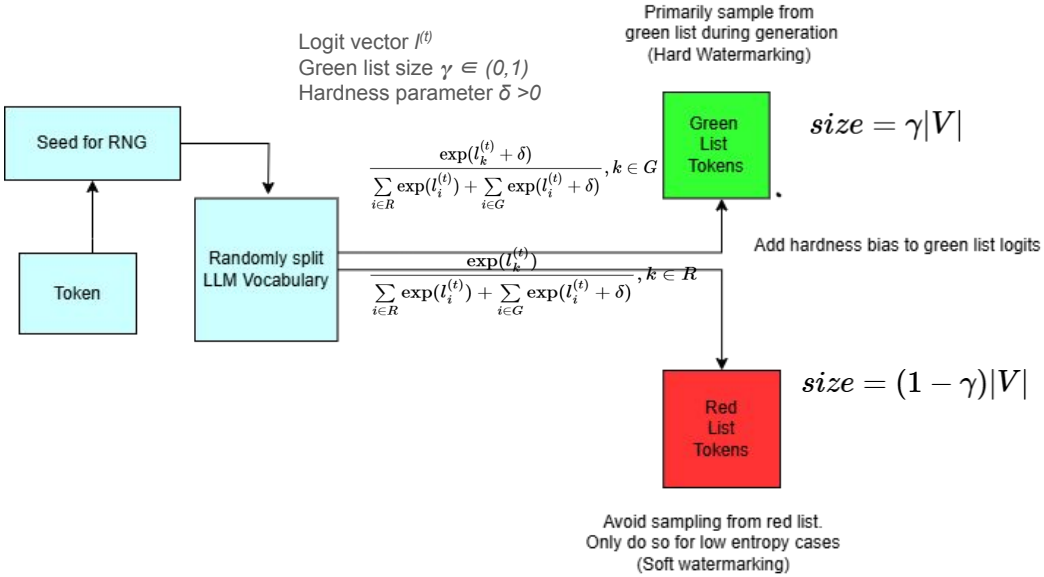


Fig: An overview of KGW Watermarking Process

Literature Review

Robust Distortion-free Watermarks for Language Models[7]

Rohith Kuditipudi, John Thickstun Tatsunori Hashimoto , Department of Computer Science Stanford University July 2023

This paper introduces **distortion-free watermarks** for language models by aligning generated text to a secret key sequence, ensuring robustness against edits while preserving the original text distribution—unlike prior hashing-based methods that bias outputs.

Key Innovation: Replaces hashing with ***sequence alignment*** for distortion-free, robust watermarks

Contribution

- **Distortion-free generation**: Watermarked text matches the LM's original distribution.
- **Prompt-agnostic detection**: No need for the LM or original prompt to verify watermarks.
- **Empirical robustness**: Withstands 40–50% token corruption (e.g., substitutions, paraphrasing).

Advantages

- No perceptual distortion [].
- Strong robustness to edits (EXP beats KGW).
- Provable guarantees via alignment cost metrics.

Disadvantages

- Slower detection (scales with key length).
- Less effective for **low-entropy** text.
- Vulnerable to **Round Trip Translation Attack** []

*A **round-trip translation attack** means translating text to another language and back to the original to see if hidden messages or patterns still remain.*

Literature Review

SEMSTAMP: A Semantic Watermark with Paraphrastic Robustness for Text Generation[9,10,11,12]

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen,Benjamin Van Durme, Daniel Khashabi, Yulia Tsvetkov in Proceedings of the 2024 Conference of the North American Chapter of ACL: Human Language Technologies

SEMSTAMP proposes a novel approach to watermarking AI-generated text by operating at the semantic level rather than the token level using **Locality Sensitive Hashing** and **contrastive learning**.Also forms custom encoder for quality assurance

Contributions:

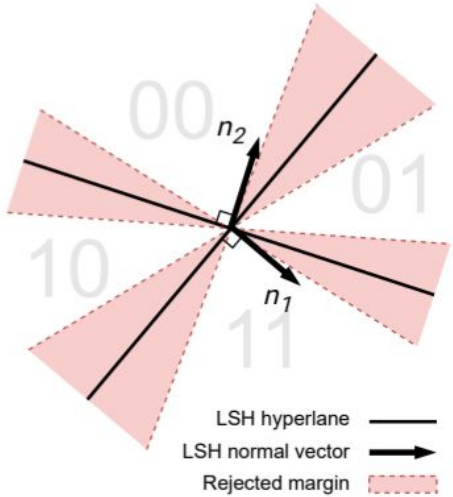
- marks whole **sentences** based on their **meaning** instead of individual words
- New **Bigram Paraphrase attack**, resistance against it and better sentence encoder with rejection margin to enforce watermark

Advantages:

- Robust against paraphrasing/token based attacks
- Maintains text quality

Drawbacks:

- Slower generation due to **rejection sampling.**
- Can be bypassed if encoder and hashing setup known



Literature Review

Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models [1]

Zhiwei He1 ,Binglin Zhou1 , Hongkun Hao1, Aiwei Liu, Xing Wang, Zhaopeng, Tu, Zhuosheng Zhang, Rui Wang in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics,2024

Investigates cross-lingual consistency in LLM watermarking and proposes **X-SIR**, a defense method that improves robustness through semantic clustering across languages.

Contributions:

- Introduced the **Cross-lingual Watermark Removal Attack** (CWRA), which effectively removes watermarks by translating prompts
- Proposed **X-SIR,** thatuses semantic clustering and robust vocabulary partitioning

Advantages:

- **X-SIR** significantly enhances watermark detectability even after translation and paraphrasing attacks.
- **X-SIR** does not degrade and can even improve the quality of generated text.

Drawbacks:

- **X-Sir** improves watermark under **CWRA** ,but is still has limitations,especially with certain language pairs(**English ⟶German**)

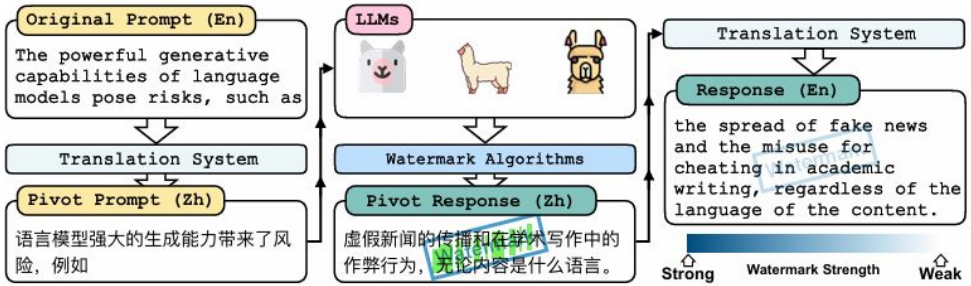


Fig:CWRA

Literature Review

Waterfall: Scalable Framework for Robust Text Watermarking and Provenance for LLMs[6]

Gregory Kang Ruey Lau, Xinyuan Niu, Hieu Dao, Jiangwei Chen, Chuan-Sheng Foo, Bryan Kian Hsiang Low, in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing(EMNLP),2024

WATERFALL, a training-free framework for watermarking text and code to protect intellectual property from LLM-based paraphrasing and unauthorized training. Its scalable design enables efficient real-world deployment without requiring model retraining.

Contributions:

- First to use LLMs as paraphrasers for watermarking.
- Waterfall, a training-free LLM-based paraphrasing framework for watermarking
- Defined the problem of scalable, robust text watermarking with real-world criteria.

Advantages:

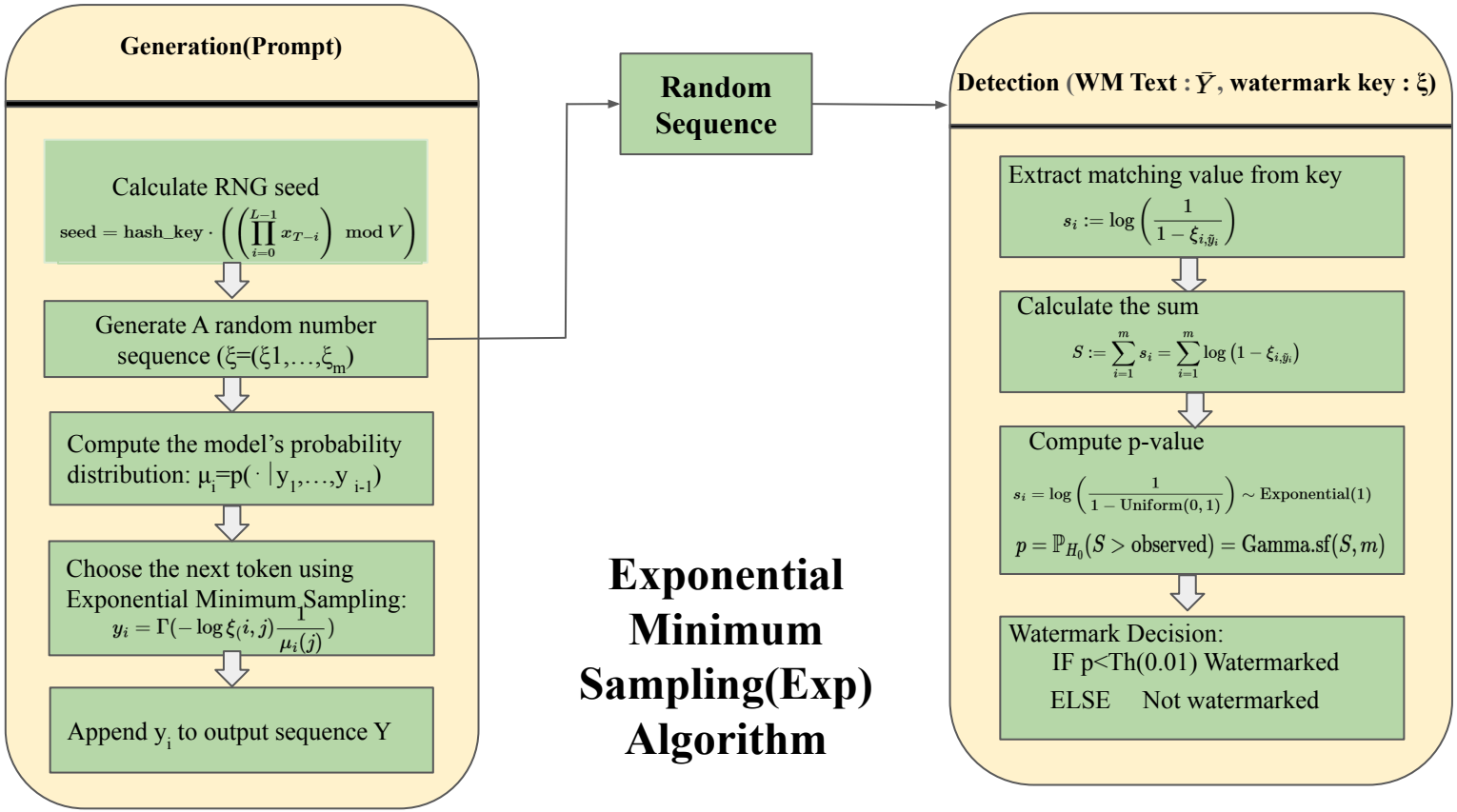
- Waterfall preserves watermark integrity under translation(**English-Spanish-English**) and paraphrasing attacks
- WATERFALL is harder for attackers to detect and remove without damaging the text.

Drawbacks:

- Less effective for short text or text with strict formatting
- For texts where style and formatting are critical (like poems or highly structured documents), paraphrasing may unintentionally alter important stylistic elements, making watermarking less effective.

Literature Review

Robust Distortion-free Watermarks for Language Models (Continue)[3]



Terminologies

Symbols used to express EXP Algorithm

Symbol	Meaning
$\mathcal{V}$	Vocabulary of the language model
$p(y \mid x)$	Probability of token y given prefix x , from the language model
$\xi \in [0,1]^N$	Random vector (watermark key), one value per token in the vocabulary.
$Y = (y_1, \ldots, y_m)$	Generated sequence of m tokens
$\mu_i := p(\cdot \mid y_1, \ldots, y_{i-1})$	Probability distribution over the next token at position i
$\Gamma(\mu_i, \xi_i)$	Sampling function (decoder) used for selecting token y_i
$\alpha(Y)$	Watermark potential of the sequence Y, measuring deviation from deterministic behavior.
x_i	Token IDs from prompt

Comparison of Watermarking Techniques

- **KGW's soft watermarking** improved imperceptibility through **green-list biasing** but remained vulnerable to **cross-lingual attacks.**

- SemaMark leveraged **semantic-level substitutions**, offering better lexical resilience but requiring **extensive synonym databases**—a challenge for Bangla.

- **EXP** algorithm introduced **distortion-free guarantees** through exponential sampling.But, quadratic complexity makes it **less scalable**

- **WATERFALL** combines vocabulary permutation and orthogonal perturbation to achieve **sublinear detection**. Notably, while **CWRA/X-SIR** specifically **address cross-lingual robustness** via semantic clustering,

- Dependency on external dictionaries **limits applicability** to languages with limited lexical resources.

Trade-off: Rule-based methods (e.g., lexical watermarking) are **lightweight but brittle**, whereas **learning-based approaches** (e.g., X-SIR) enhance **robustness** at the cost of **increased complexity.**

Gaps and Opportunities in Current Research

Key Gaps:

• **Language Disparity:** Most methods (e.g., KGW, WATERFALL) focus on English, neglecting Bangla's agglutinative morphology and low-resource constraints.

• **Attack Resilience:** Cross-lingual attacks (e.g., CWRA) and paraphrasing remain understudied in non-Latin scripts.

• **Evaluation Metrics:** Existing frameworks lack standardized benchmarks for semantic preservation in low-entropy Bangla text (e.g., proverbs).

Emerging Opportunities

• **Hybrid Embedding:** Combining token-based biasing (KGW) with semantic clustering (SEMSTAMP) or post-processing algorithms(like , WATERFALL) could enhance robustness against translation attacks.

• **Dynamic Entropy Adaptation:** Algorithms could prioritize high-entropy segments in Bangla (e.g., creative writing) while preserving low-entropy content (e.g.,formal documents).

• **Resource-Efficient Training:** Leveraging multilingual LLMs (e.g., Llama -3 8B) to bootstrap watermarking for Bangla with minimal labeled data.

Proposed Methodology

The primary goal of this research

- **Watermark Embedding Mechanism**: We apply two parallel watermarking techniques — **KGW Soft Biasing** and **Exponential Sampling** during Bangla text generation using a bangla instruction-tuned **LLaMA-3-8B** model.
- **Round-Trip Translation Attack Simulation**: The watermarked text undergoes Bangla → English → Bangla translation to simulate RTT attacks and test watermark robustness.
- **Watermark Detection and Robustness Evaluation**: We use z-score analysis to detect watermarks pre- and post-attack, and assess semantic quality and perplexity shifts using ROUGE metrics.
- **Combining Pre-Processing with Post-Processing(Future Work)**: We aim to enhance robustness by integrating **SEMSTAMP** and **WATERFALL** watermarking methods.

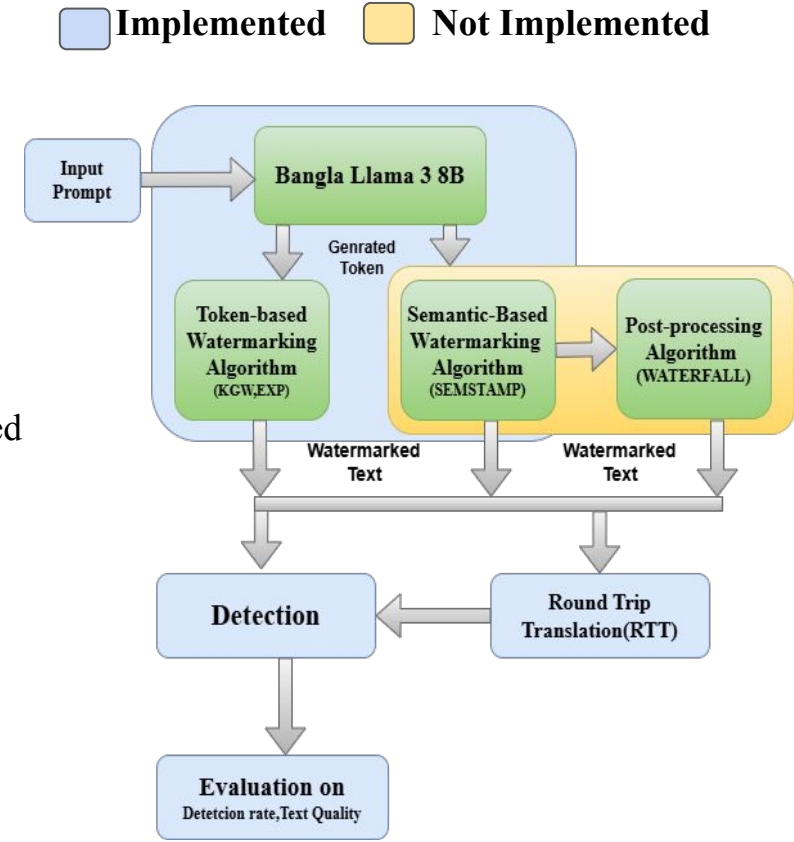


Figure: Workflow Interconnection

Datasets and Experimental Setup

Dataset

We used a subset of samples from ***BanglaLLM/bangla-alpaca-orca instruct*** dataset to evaluate .Our custom dataset structure is shown below."

Instruction(Prompt)	Generated Output (Watermarked)	Generated Output (Not Watermarked)	IS Watermarked? (0 or 1)

Experimental Setup

Model Configuration

- **Model:** *BanglaLLM/BanglaLLama-3-8b-bangla-alpaca-orca instruct-v0.0.1*
- **Quantization:** 8-bit (optimized for efficiency)

Watermarking Parameters

- **Algorithm:** KGW , EXP watermarking
- **Key Hyperparameters:**
  - max_new_tokens: 500
  - min_length: 50
  - do_sample: True (enables probabilistic sampling)
  - no_repeat_ngram_size: 5 (prevents repetitive outputs)

Evaluation Metrics

- **Detection Accuracy:** Ratio of correct watermark detections
- **Perplexity:** Fluency measure (lower is better)
- **ROUGE Scores:** ROUGE-1, ROUGE-2,ROUGE_-l
- **Robustness Tests:**
  - Paraphrasing attacks
  - Round-trip translation (Bengali → English → Bengali)

BanglaLLM/BanglaLLama-3-8b-bangla-alpaca-orca instruct-v0.0.1

Results

Sample output example

Figure 4.1: Watermarked Results for KGW and EXP

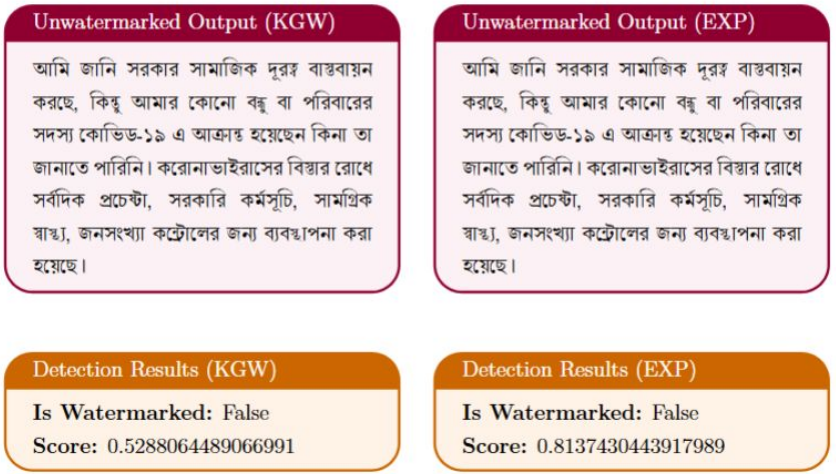


Figure 4.2: Unwatermarked Results for KGW and EXP

Results

A Quantitative analysis between two algorithm we experimented and evaluated

Criteria	KGW	EXP
Detection Accuracy	0.885	0.912
Average Watermarked Perplexity (↓ better)	2.309	2.0295
Average Unwatermarked Perplexity (↓ better)	2.1616	2.1368
ROUGE-1 (↑ better)	0.3670993	0.3853785
ROUGE-2 (↑ better)	0.3038986	0.3206552
ROUGE-L (↑ better)	0.3600586	0.3783677
Robustness to Round-Trip Translation (↑ better) (detection rate)	0.09	0.13

Results Analysis

EXP method showing marginally superior performance in detection accuracy and text fluency

Divergence from Hypotheses: The near-identical outputs for **KGW** and **EXP** in some cases were unexpected in ***Figure 4.1*** This could arise from the model's strong preference for certain high-probability Bengali phrases, limiting the watermark's variability.Future work could explore entropy-aware prompting to mitigate this.

- **Detection Accuracy**:
  EXP achieves higher accuracy, offering better resistance to false negatives. KGW is slightly less accurate but remains robust, consistent with earlier English LLM studies.
- **Perplexity and Fluency**:
  Both methods maintain similar perplexity ***(~2.0–2.3)*** to unwatermarked text ***(~2.1)***. EXP shows slightly better fluency due to dynamic thresholding, while KGW's fixed greenlist may penalize rare tokens.
- **Semantic Preservation (ROUGE Scores)**:
  EXP yields higher ROUGE-L ***(0.378 vs. 0.360)***, preserving long-sequence meaning better. KGW shows a slight drop in ROUGE-2 ***(0.303 vs. 0.320)***, indicating minor bigram disruption.
- **Robustness Challenges**:
  Both methods are vulnerable to round-trip translation attacks ***(detection scores: 0.09–0.13)***, consistent with known challenges in multilingual watermarking.

Limitations And Challenges

Despite its contributions, this study has several limitations:

- **Dataset Size and Diversity:** The lack of expansive, varied Bangla datasets constrained the evaluation of watermark robustness across different domains.
- **Computational Overhead :** Semantic-aware rejection sampling introduced additional latency in text generation.
- **Adversarial Weaknesses:** While robustness was enhanced, paraphrase-based and generative attacks still pose threats.
- **Language-Specific Trade-offs:** Preserving text quality while embedding detectable watermarks remained more challenging in complex, low-entropy Bangla text.

There was some challenges we faced:

- Lack of prior studies in Bangla watermarking
- Experimental setup issues
- Evaluation difficulty
- Trade-off between detection strength and text naturalness

Summary of Findings

Through critical analysis and experimentation, several important findings emerged:

- ➢ Direct application of English-based watermarking methods such as KGW and EXP to Bangla was insufficient, highlighting the need for language-specific adaptation.
- ➢ Our Bangla-focused watermarking approach demonstrated stronger resistance to adversarial attacks while preserving text semantics.
- ➢ The flexible and rich morphology of Bangla posed unique challenges for embedding imperceptible watermarks, which were addressed through careful algorithmic design.
- ➢ Empirical evaluations confirmed that watermark detectability and resilience improved significantly when tailored strategies were employed for Bangla

Future Directions

The lack of large, diverse Bangla datasets limited comprehensive robustness evaluation. Developing fully resilient watermarking algorithms against round-trip translation attacks for Bangla LLMs remains an open research challenge.

References

Works cited in this presentation

[1] Z. He, B. Zhou, H. Hao, A. Liu, X. Wang, Z. Tu, Z. Zhang, and R. Wang, "Can watermarks survive translation? On the cross-lingual consistency of text watermark for large language models," *arXiv preprint* arXiv:2402.14007, 2024.

[2] J. Achiam, S. Adler, S. Agarwal, et al., "GPT-4 technical report," *arXiv preprint* arXiv:2303.08774, 2023.

[3] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR, 2023, pp. 17061–17084.

[4] S. J. Chanchani and R. Huang, "Composition-contrastive learning for sentence embeddings," *arXiv preprint* arXiv:2307.07380, 2023.

[5] H. Touvron, L. Martin, K. Stone, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint* arXiv:2307.09288, 2023.

[6] G. K. R. Lau, X. Niu, H. Dao, J. Chen, C.-S. Foo, and B. K. H. Low, "Waterfall: Scalable framework for robust text watermarking and provenance for LLMs," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 20432–20466.

[7] R. Kuditipudi, J. Thickstun, T. Hashimoto, and P. Liang, "Robust distortion-free watermarks for language models," *arXiv preprint* arXiv:2307.15593, 2023.

[8] X. He, Q. Xu, L. Lyu, F. Wu, and C. Wang, "Protecting intellectual property of language generation APIs with lexical watermark," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10758–10766.

[9] K. Yoo, W. Ahn, J. Jang, and N. Kwak, "Robust multi-bit natural language watermarking through invariant features," *arXiv preprint* arXiv:2305.01904, 2023.

References

[10] Y. Fu, D. Xiong, and Y. Dong, "Watermarking conditional text generation for AI detection: Unveiling challenges and a semantic-aware watermark remedy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 18003–18011.

[11] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC)*, 1998, pp. 604–613.

[12] J. Wieting, K. Gimpel, G. Neubig, and T. Berg-Kirkpatrick, "Paraphrastic representations at scale," *arXiv preprint* arXiv:2104.15114, 2021.