REMARK-LLM: A Robust and Efficient Watermarking Framework for Generative Large Language Models

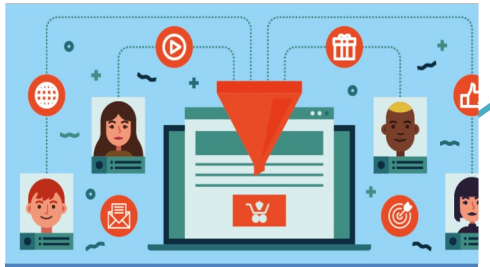
Ruisi Zhang, Shehzeen Samarah Hussain, Paarth Neekhara, Farinaz Koushanfar

UC San Diego

Large Language Model Applications

Health Care

Software Develop



Content Generation

Translation

Search Engines

Education

Motivation

- The White House published an executive order for safe Generative AI; Watermarking is highlighted in it for authenticating and detecting LLM generated content.

(gg) The term "watermarking" means the act of embedding information, which is typically difficult to remove, into outputs created by AI — including into outputs such as photos, videos, audio clips, or text — for the purposes of verifying the authenticity of the output or the identity or characteristics of its provenance, modifications, or conveyance.

Sec. 4. Ensuring the Safety and Security of AI Technology.

(i) authenticating content and tracking its provenance;

(ii) labeling synthetic content, such as using watermarking;

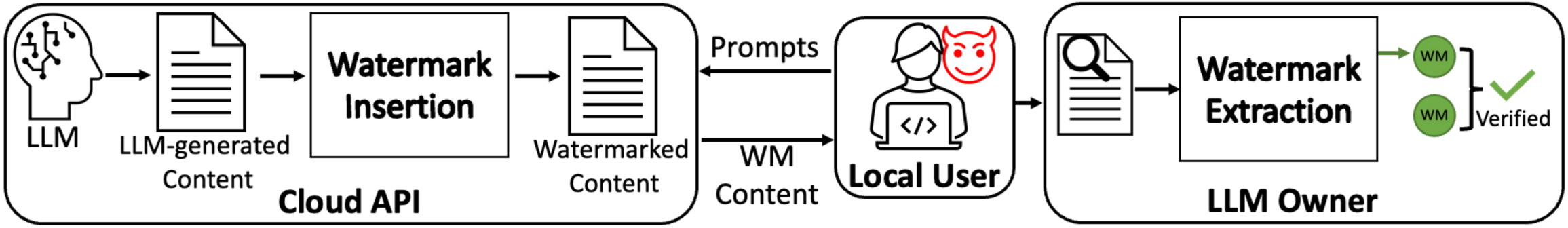
(iii) detecting synthetic content;

(iv) preventing generative AI from producing child sexual abuse

(C) reasonable steps to watermark or otherwise label output from generative AI;

Watermarking Global Flow

- Watermark Insertion
  - The LLM-generated content is watermarked with owner's signature before sending to local user.
- Watermark Extraction
  - The LLM owner claims his ownership by decoding the signature from the watermarked content.

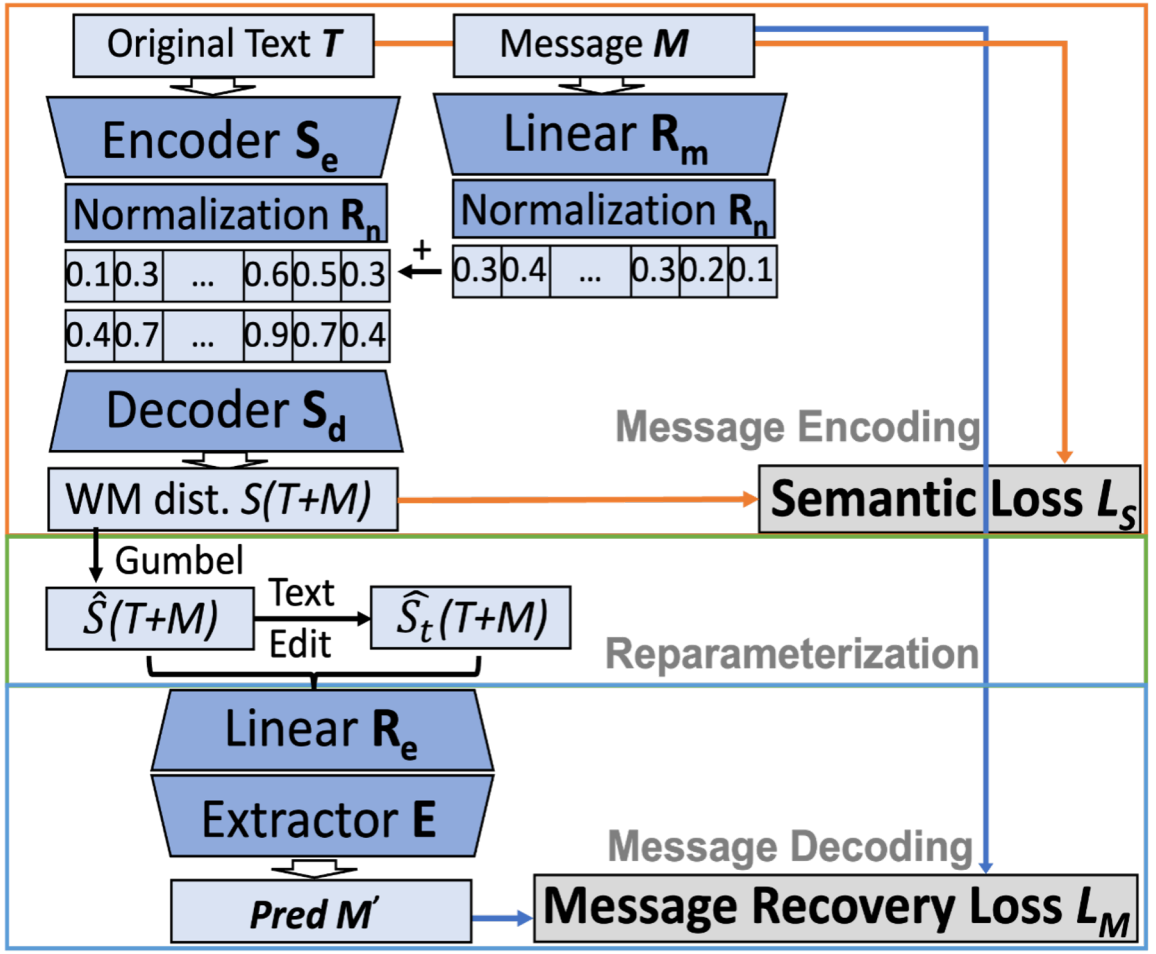


Challenges for Text Watermarking

- Sparsity
  - LLM-generated texts usually have a few thousand tokens.
  - 256×256 pixel Images have 65k potential pixels for wm insertion
- Sensitivity
  - Minor changes in texts' tokens can distort its meanings and fluency.
  - Such changes in images are more imperceptible.
- Vulnerability

REMARK-LLM Overview

- Watermark Insertion
  - The backbone is a Seq2Seq model (T5) takes original LLM-generated text and signature as input, and generates watermarked text
- Watermark Extraction
  - The transformer-based extractor decodes watermark signatures from the watermarked texts & malicious transformed wm texts.



End-to-End Training

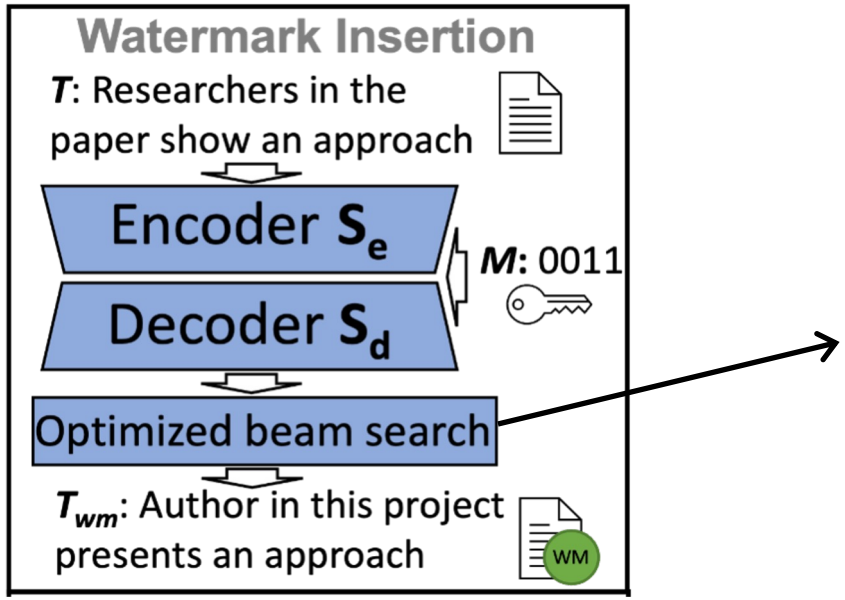
- Training Objectives
  - Semantic Loss
    - Minimize the cross-entropy loss between input texts and the watermarked texts distributions.
  - Message Recovery Loss
    - Minimize the L1 loss between input message and decoded message
    - Minimize the L1 loss between input message and malicious transformations' decoded messages.

$$L_{total} = L_{CE}\big(T, S(T+M)\big) + L_{l1}(M, M') + L_{l1}(M, M'_t)$$

Semantic Loss Message Recovery Loss

Watermark Insertion

- The watermark insertion module leverages an optimized beam search module to decode readable texts from the watermarked distribution $S(T + M)$.



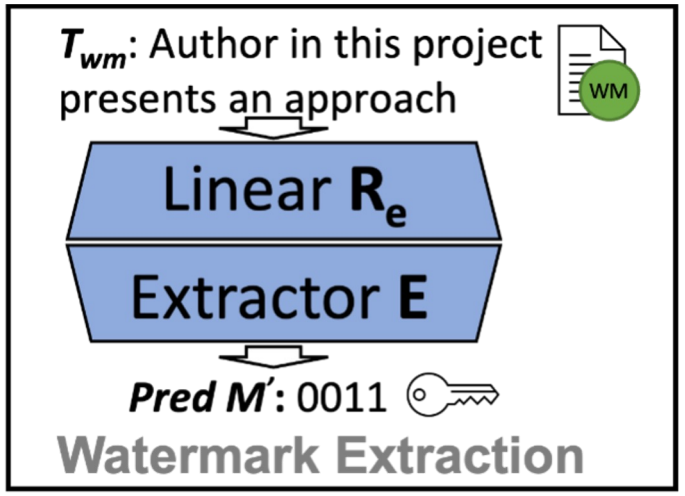
Algorithm 1 Optimized Beam Search Algorithm

Require: LLM-generated text token $\bar{T}$, temperature list τ, beam size B, number of iterations K, message $\bar{M}$
Ensure: Watermarked text $\bar{T}_{wm}$

Initialize max_accuracy = 0
Initialize $\bar{T}_{wm}$ = None
for k = 1 to K **do**
    Initialize mask $\bar{T}_M$
    Initialize watermarked dist. $\mathcal{S}(\bar{T} \cdot \bar{T}_M + \bar{M})$
    for each $\mathcal{S}_i$ in $\mathcal{S}(\bar{T} \cdot \bar{T}_M + \bar{M})$ **do**
        $\mathcal{S}_{\text{noisy},i} \leftarrow \mathcal{S}_i + \text{Gumbel}(\mathcal{S}_i, \tau_k)$
    end for
    $T_k \leftarrow \text{Beam\_Search}(\mathcal{S}_{\text{noisy}}, B)$
    for each T_{ki} in T_k **do**
        $a \leftarrow \text{Accuracy}(\mathbf{E}(T_{ki}), \bar{M})$
        if $a >$ max_accuracy **then**
            max_accuracy $\leftarrow a$
            $\bar{T}_{wm} \leftarrow T_{ki}$
        end if
    end for
end for
return $\bar{T}_{wm}$

Watermark Extraction

- The messages are decoded via watermark extraction module.
- The encoded messages M and decoded M' are compared to calculate watermarking strength (z-score) and claim ownership.



N-bit matches between M and M'

$$Z = \frac{|N| - \mu}{\sigma}$$

Variance: $\sigma^2 = |M| \times p \times (1 - p)$

Mean: $\mu = |M| \times p$

Experiment Setup

- Metrics
  - **BERT Score**: Semantic preservation between the original and watermarked texts.
  - **BLUE-4**: The coherence between the original and watermarked texts.
  - **WER**: The percentage of decoded message matches inserted ones
- Baselines
  - KGW & EXP: Inference-based watermarking
  - AWT: Neural-based watermarking
  - CATER: Rule-based watermarking

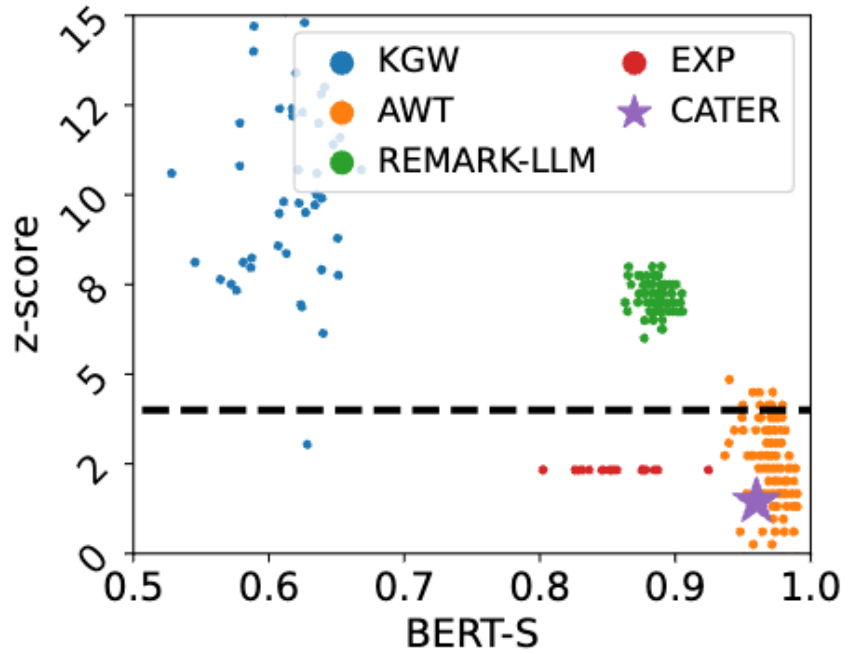
Long Sequence Watermarking (640 token)

- **High Watermark Extraction Rates**: 95%+ WER when decoding 64 bit signatures from 640 token texts.
- **High Semantic preservation**: ~0.9 BERT-S between watermarked and original texts.
- **High transferability**: REMARK-LLM is trained on HC3, and successfully watermark three new datasets without additional fine-tuning.

Dataset	Methods	64 bits WER(%) ↑	64 bits BERT-S ↑	64 bits BLUE-4 ↑
HC3	REMARK-LLM	95.61	0.91	0.41
	AWT [1]	68.42	**0.95**	**0.82**
	KGW [15]	**99.57**	0.58	0.01
	EXP [17]	64.68	0.80	0.01
	CATER [11]	81.25	0.65	0.30
WikiText-2	REMARK-LLM	94.48	0.85	0.16
	AWT [1]	65.77	**0.96**	**0.85**
	KGW [15]	**99.13**	0.61	0.02
	EXP [17]	64.68	0.82	0.01
	CATER [11]	81.25	0.65	0.30
ChatGPT Abstract	REMARK-LLM	95.04	0.89	0.27
	AWT [1]	62.39	**0.95**	**0.84**
	KGW [15]	**99.01**	0.61	0.01
	EXP [17]	64.68	0.80	0.01
	CATER [11]	81.25	0.65	0.30
Human Abstract	REMARK-LLM	95.39	0.87	0.15
	AWT [1]	63.52	**0.94**	**0.85**
	KGW [15]	**98.79**	0.69	0.01
	EXP [17]	64.68	0.81	0.01
	CATER [11]	81.25	0.65	0.30

Watermarking Strength

- REMARK-LLM successfully provides strong signature insertions without compromising the watermarked texts' semantics.



Prior work demonstrates very sensitive trade-offs between semantic preservation (BERT-S) and watermarking strength (z-score)!

Watermarking Transferability

- REMARK-LLM is agnostic to texts generated by different LLM architectures or prompts sources.
  - We select 2k instruction prompts from Alpaca Dataset, and watermark the outputs generated by different LLMs.

Model	WER(%)	BERT-S	BLUE-4
OPT-2.7B	93.42	0.91	0.34
OpenOrca-7B	93.70	0.92	0.35
LLaMA-2-7B	91.18	0.91	0.39

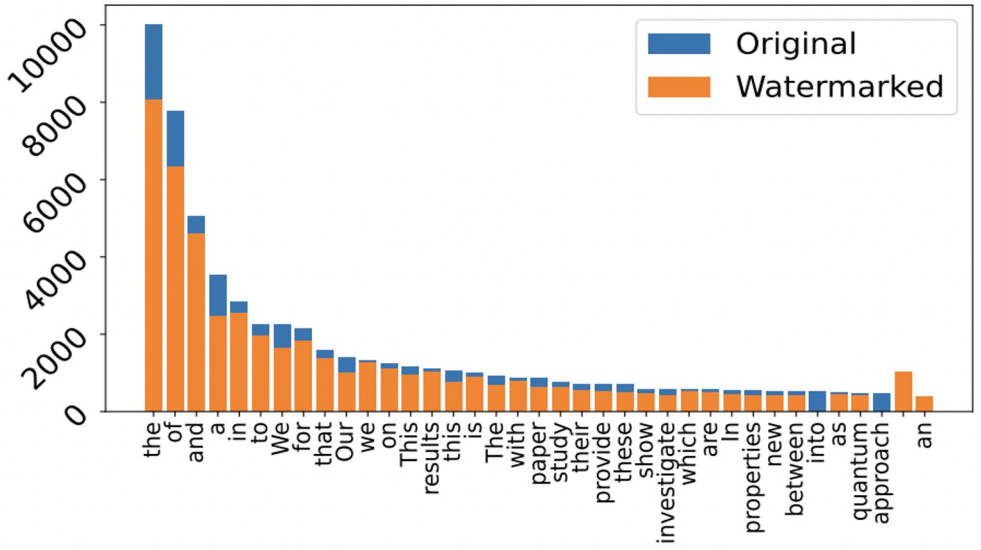
Watermarking Examples

- REMARK-LLM learns to (1) replace the words with their synonyms and (2) edit contents to ensure coherence.

Original Text	Watermarked Text
It can be hard to explain it in simple terms. But I'll do my best! During inflation, the universe expands at an incredibly fast rate. But it's important to note that this expansion is not like the movement of objects through space.	It can be hard directly explain exactly in simple terms. But I'll try my best! During time, the universe expands at an infinite incredibly fast rate. But it's important to understand that this expansion is not about the movement of objects through space itself.
In the context of financial investments, "headwinds" refer to negative factors that can potentially hinder the performance of an investment. These may include economic conditions, regulatory changes, market trends, or other external factors that can work against the investment.	In the context of stocks investing, "headwinds" refer for negative factors that can potentially impact the value of an investment. These factors include economic impacts, regulatory issues, market conditions, and other external factors that work against the investment.
The paper discusses Colombeau's generalized function on arbitrary manifolds. We first define the space of Colombeau's generalized functions by quotienting out by a suitable ideal endowed with a ring structure.	The paper introduced Colombeau's generalization function on arbitrary manifolds. We first study the space of Colombeau's generalized functions by particle out ideal endowed with a ring structure.
This paper presents a novel methodology for constructing super throats using non trivial scalar fields. By introducing these fields, we are able to achieve unprecedented control over the dynamics of the throats.	This research presents a novel approach for constructing super throats through non trivial scalar fields. By utilizing these fields, we are able to obtain precise control over the dynamics of the throats.

Watermark Detection Attack

- By comparing the original and watermarked texts, the adversarial cannot detect if the content is watermarked or not.



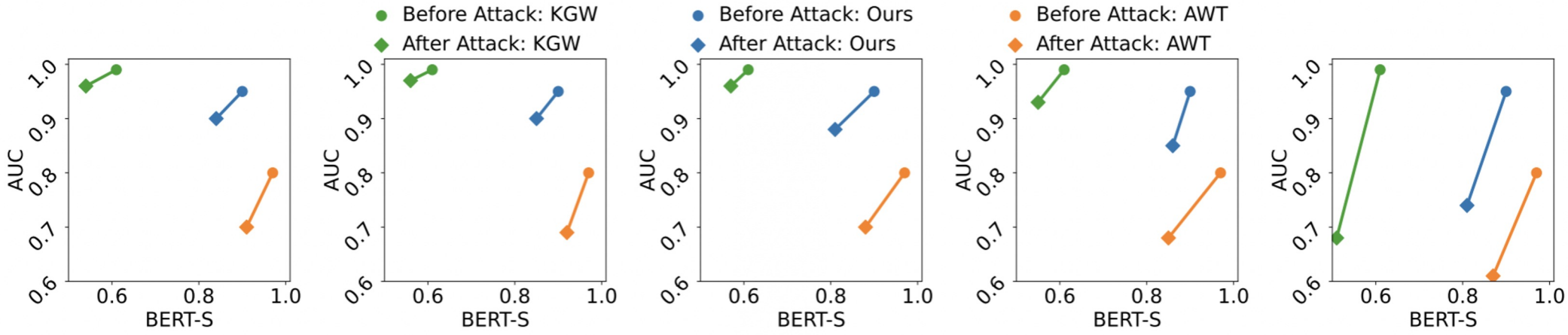
Detection with word distribution analysis

Model	Acc. (%)	F1-Score
Transformer [36]	50.00	0
BERT-base [8]	50.00	0
BERT-large [8]	50.00	0

Detection with machine learning-based models

Watermark Removal Attack

- REMARK-LLM maintains high AUC under text edits, rephrase, and re-watermark attacks.
  - The robustness is maintained all while preserving the watermarking fidelity.



(a) Text Deletion Attack (b) Text Addition Attack (c) Text Replacement Attack (d) Text Rephrase Attack (e) Re-watermark Attack

Conclusions

- Introduction of REMARK-LLM, a watermarking framework that has **high capacity**, **high transferability**, and **robustness**.
- Extensive evaluations on various LLM-generated and human-written benchmarks demonstrate the effectiveness of REMARK-LLM.
- Checkout our paper & code 👇

  Paper: https://arxiv.org/abs/2310.12362
  Code: https://github.com/ruisizhang123/REMARK-LLM