



# An End-to-End Model for Logits-Based Large Language Models Watermarking

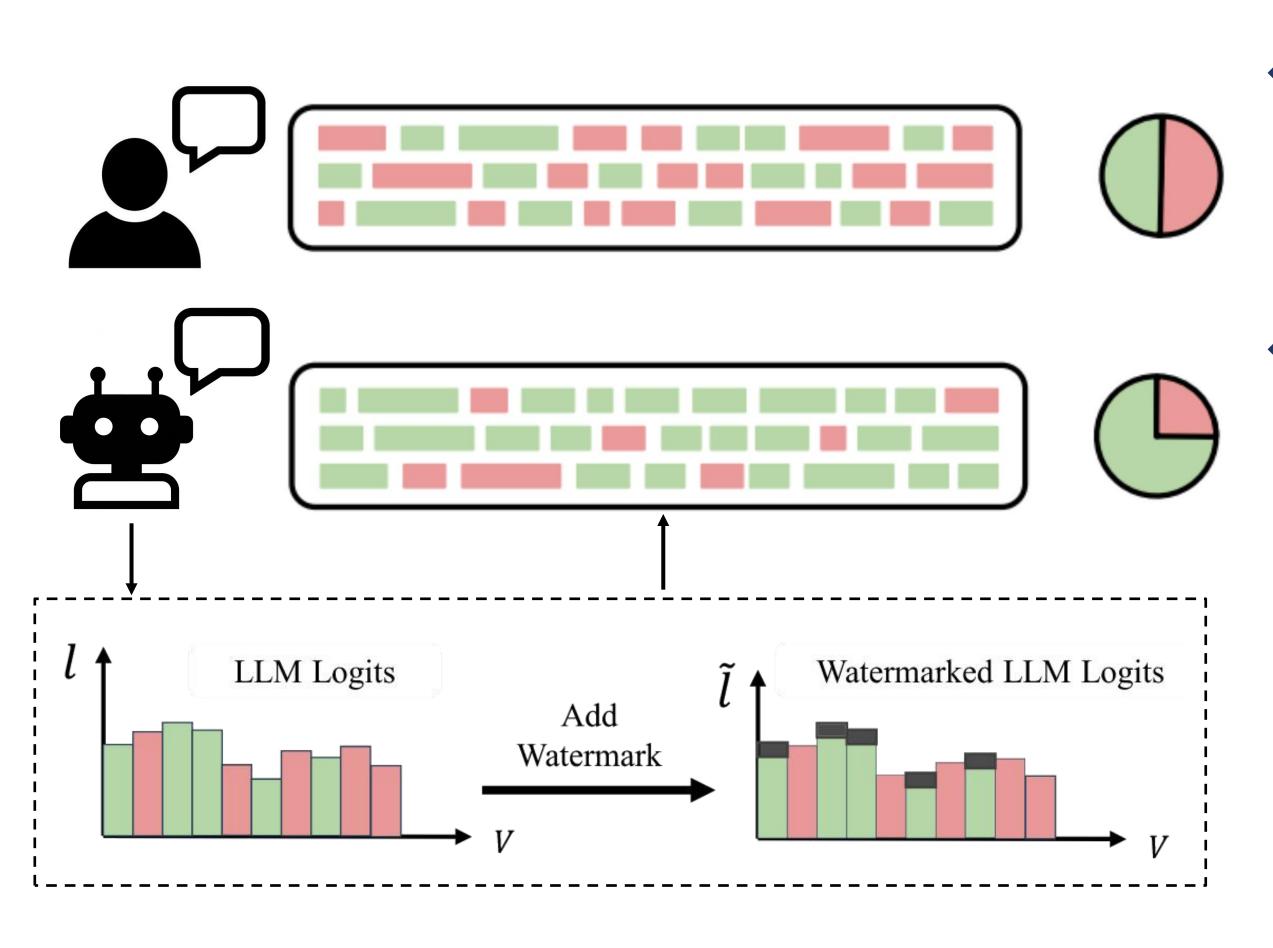
Kahim Wong, Jicheng Zhou, Jiantao Zhou, Yain-Whar Si







## Logits-Based LLM Watermarking

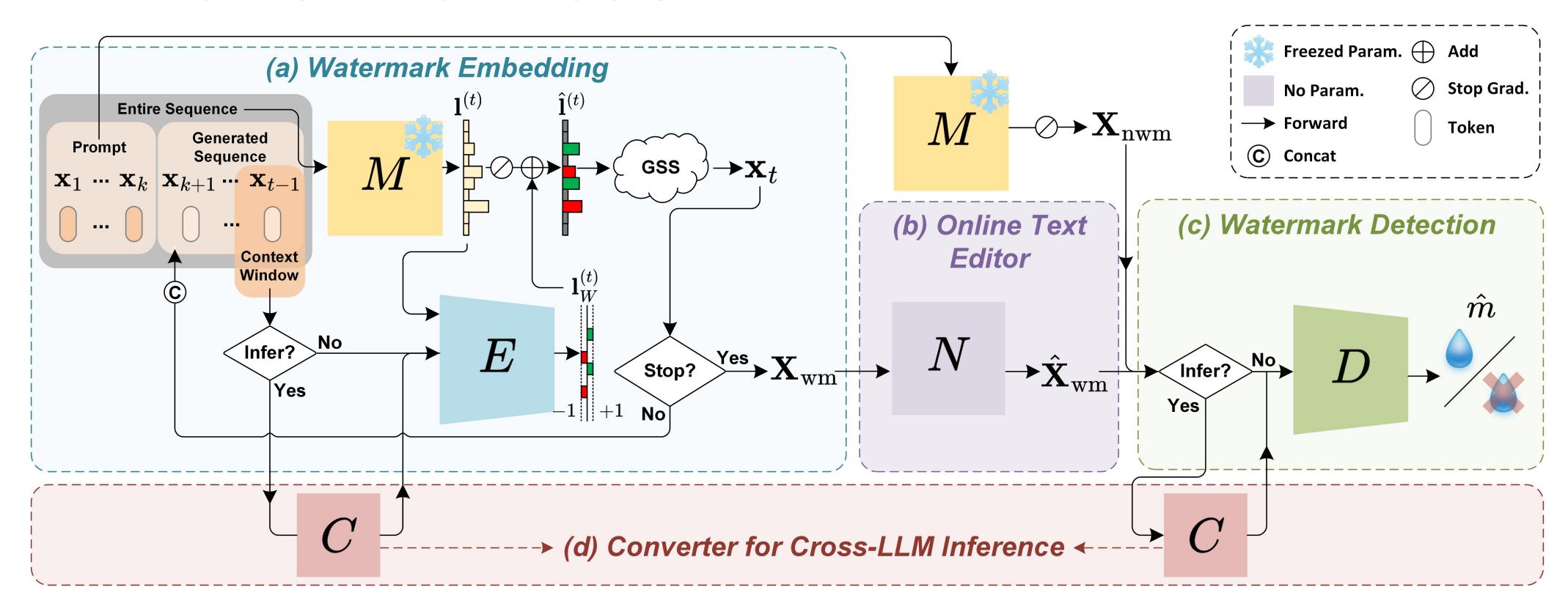


- Passive detectors (e.g. DetectGPT)
   face high false positives when
   identifing human from AI text
- Watermark is more reliable:
- 1. Hash context to randomly split vocab into green and red sets
- 2. Add bias to raises probability of green tokens during generation
- 3. Flag text has high proportion of green tokens as Al-generated

## Challenge

- High accuracy on clean watermarked text, but performance drops once the text is edited (e.g. paraphrasing)
- ◆ The added bias can harm the LLM performance on downstream tasks

### End-to-End Model



- End-to-End training: Encoder (add bias), Text Editor (simulate edits), and Decoder (detect watermark) are jointly trained to maximize detection accuracy and preserving LLM output quality
- Online prompting: Dynamically prompt the on-the-fly LLM to transforms non-differentiable operations (e.g. online paraphrasing and semantic computation) into differentiable
- ◆ Converter enable cross-LLM inference without retraining

## Experiments

- ◆ Evaluate on MarkLLM benchmark: we train exclusively on *OPT-1.3B* and use the converter at inference.
- Our method deliver stronger watermark robustness while maintaining output quality

#### CL: Clean sample; SS: Synonymous substitution; CP: Copy-paste; PA: Paragraphing; PPL: perplexity

<u> </u>	OPT-1.3B					Llama2-7B					Qwen2.5-7B				
Method	ethod Robustness (F1 <sup>†</sup> )			Quality	Robustness (F1↑)			Quality	Robustness (F1↑)		Quality				
	CL	SS	CP	PA	PPL↓	CL	SS	CP	PA	PPL↓	CL	SS	CP	PA	PPL↓
NWM	-	-	<del>-</del> 1	0.450 0. <del></del>	10.484	-	-			6.811	0-		-	-	8.921
KGW	1.000	0.990	0.983	0.880	13.173	1.000	0.970	0.846	0.858	8.658	1.000	0.983	0.975	0.832	11.419
Unigram	1.000	0.997	0.943	0.943	12.739	0.995	0.990	0.873	0.909	9.275	1.000	0.993	0.955	0.942	10.847
Unbiased	0.992	0.800	0.949	0.680	11.940	0.990	0.785	0.912	0.684	7.565	0.985	0.780	0.930	0.683	10.061
DiPmark	0.997	0.809	0.954	0.692	12.085	0.983	0.779	0.915	0.670	7.681	0.985	0.780	0.923	0.681	10.488
Ours	0.998	0.992	0.975	0.952	12.397	0.995	0.985	0.978	0.916	7.730	0.995	0.985	0.985	0.945	9.997
$\Delta_{ ext{Unigram}}$	0%	-1%	+3%	+1%	+3%	0%	0%	+12%	+1%	+17%	-1%	-1%	+3%	0%	+8%
$\Delta_{ m DiPmark}$	0%	+23%	+2%	+38%	-3%	+1%	+26%	+7%	+37%	-1%	+1%	+26%	+7%	+39%	+5%

		More	e LLM	S			
	F	Robustne	Quality				
LLM	CL	SS	CP	PA	PPL↓		
		Oı	Ours	NWM			
Mixtral-7B	0.990	0.970	0.987	0.916	10.219	8.711	
Llama3-8B	0.998	0.990	0.990	0.934	7.256	5.964	
<i>Llama3.2-3B</i>	0.997	0.995	0.993	0.947	7.599	6.301	

		Dowi	<u>nstrear</u>	<u>n tasks</u>						
Metric	NWM	KGW	Unigram	Unbiased	DiPmark	Ours				
Machine translation task with NLLB-600M										
BLEU↑	31.789	26.325	26.057	28.949	28.942	31.062				
Code generation task with Starcoder										
pass@1↑	43.0	22.0	33.0	36.0	36.0	34.0				

#### Watermark Example

ahead, after the latest effort to keep the government running was knocked off course. Congress has three weeks to reach an agreement, or federal agencies will start to run out of money and stop functioning. "I think we are close to an agreement that will make sure the government continues to operate, but we are going to have a number of issues to deal with," Lew told ABC's "This Week" program in an interview that aired on Sunday. The debt ceiling, which governs the amount of money the federal government is allowed to borrow, is also set to be reached on Thursday.

Democratic Senator Chuck Schumer, the chamber's No. 3 Democrat, told CBS's "Face the Nation" program that lawmakers could avert a debt crisis if Republicans would agree to raise the borrowing cap by at least \$1 trillion. "Our position is pretty simple

WASHINGTON (Reuters) - U.S. Treasury Secretary Jack Lew on Sunday warned Congress against manufacturing a crisis over federal spending in the months ahead, after lawmakers on both sides of the political aisle said they will not back down from their demands for fiscal concessions. In the latest of a long-running game of political chicken, congressional Republican leaders said their top fiscal-policy priority is to use the need to lift the U.S \$16.7 trillion federal debt ceiling to extract spending cuts, slash tax rates and widen the fiscal gulf with President Barack Obama. "I'm not optimistic that the Congress is ready to make the tough decisions that we have to make to get our spending in control, to make the decisions to start to solve the problem of \$17 trillions in national debt that's just untenable for our economy and untenable over the long-run for our national well-being and our standard of living," Lew said in an interview on

