Watermark Smoothing Attacks against Language Models

Hongyan Chang, Hamed Hassani, Reza Shokri

NUS National University of Singapore

Penn University of Pennsylvania

Watermark in LLMs

- Detecting AI-generated content is crucial for combating fraud, misinformation, and plagiarism
- LLM Watermarking: Embeds subtle, detectable patterns in generated text while preserving quality

Watermark Effectiveness Analysis

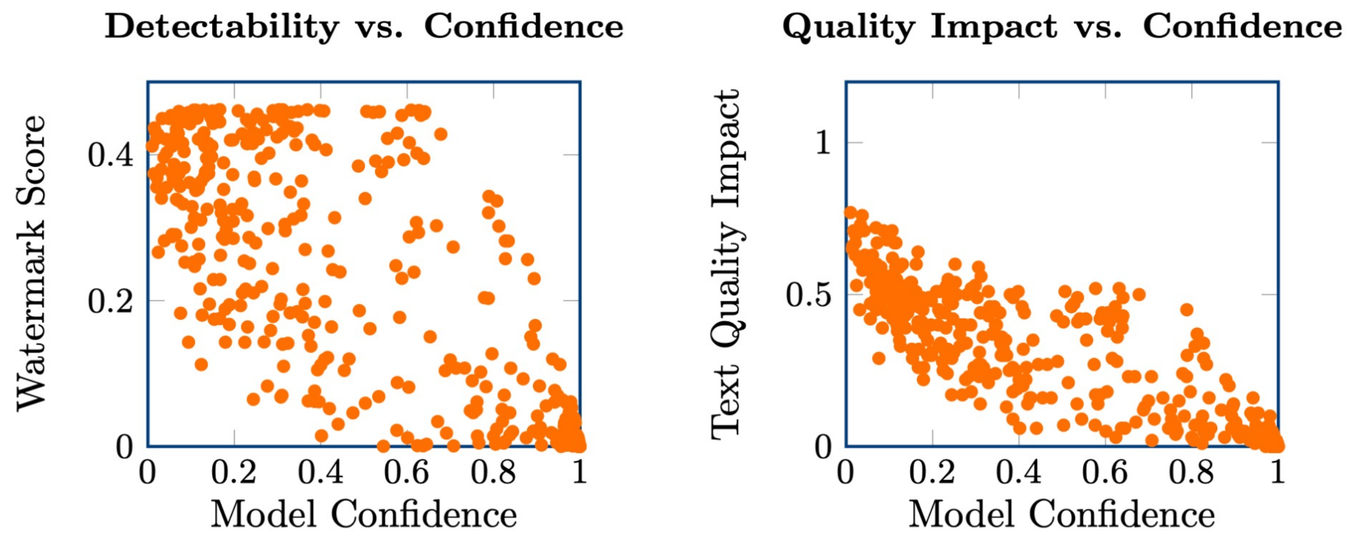
Q: Can LLM Watermarks Achieve Both High Text Quality and High Detectability?
A: Yes. We analyze how ***each token*** prediction affects watermark detection and deviates from the behavior of an unwatermarked model.

When model is confident:

- Watermarking has minimal influence
- Watermark traces are subtle
- Text quality remains high

When model is *not* confident:

- Watermarking distorts token choice
- Watermark becomes easier to detect
- Text quality is noticeably reduced



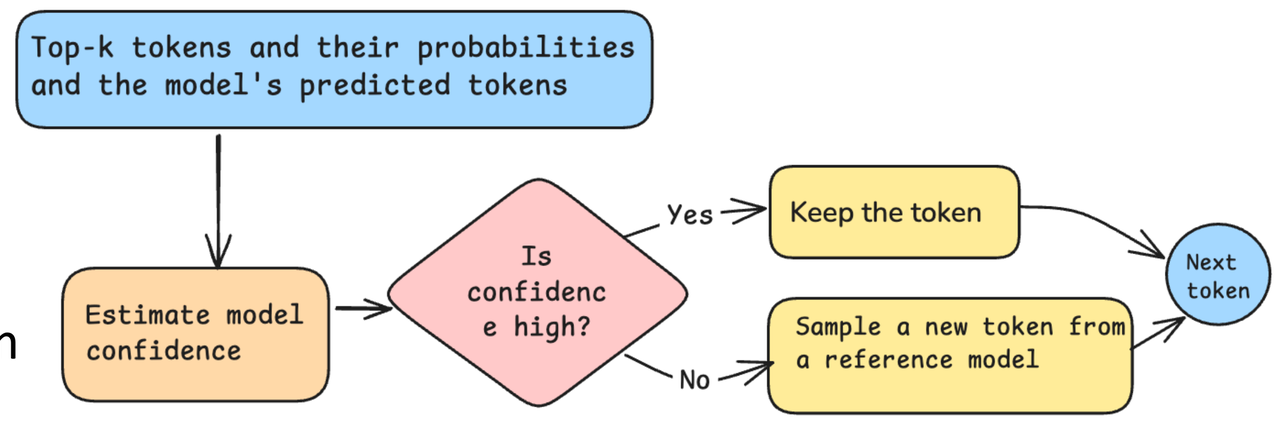
Left: Lower model confidence leads to higher watermark detectability. Right: Lower model confidence leads to more text quality degradation from watermarking.

Key Finding: a fundamental trade-off between watermark detectability and text quality
Tokens that increase watermark visibility also degrade the naturalness of the generated text — making it impossible to optimize for both at the same time.

Smoothing Attack

Core idea: Use model confidence to selectively replace tokens with high watermark score contribution while maintaining text quality

Attack Knowledge: Only needs access to top-K tokens and their probabilities (K=1, 5, or 10) from the watermarked model and a reference model



Empirical Results

Model	Attack	True Positive Rate (TPR) %								Perplexity (PPL)							
		KGW	Unigram	SynthID	DIP	Unbiased	UPV	EWD	SWEET	KGW	Unigram	SynthID	DIP	Unbiased	UPV	EWD	SWEET
OPT-1.3B	Watermarked	100	100	100	100	100	99	100	100	14.61	14.99	7.12	13.73	13.61	11.65	15.23	14.36
	Paraphrasing	3	53	1	0	3	34	0	0	14.82	14.51	10.57	13.95	14.45	13.73	14.95	14.57
	Smoothing	0	5	0	6	27	20	0	0	9.57	9.44	10.40	9.34	9.19	10.01	9.93	9.59
Llama3-8B	Watermarked	99	99	99	84	84	83	100	99	4.60	4.61	4.83	4.03	4.02	4.38	4.56	4.53
	Paraphrasing	2	54	1	0	2	2	7	14	5.35	5.60	5.62	5.25	5.36	5.43	5.73	5.64
	Smoothing	2	24	0	6	5	1	3	4	3.20	3.10	3.40	3.17	3.17	3.12	3.13	3.09
Qwen2-1.5B	Watermarked	100	100	100	100	100	86	100	100	16.46	15.41	6.94	14.34	14.64	11.93	16.31	15.89
	Paraphrasing	2	5	1	2	1	2	1	4	10.45	10.40	6.90	10.10	9.97	9.03	10.18	10.18
	Smoothing	0	1	0	11	5	0	0	0	8.02	7.77	10.21	7.62	7.68	8.16	7.82	7.85

TPR: True Positive Rate (lower is better), PPL: Perplexity (lower is better).

- Our attack completely removes watermarks in many cases (TPR reduced to 0%).
- Our attack even outperforms paraphrasing attack with GPT-3.5-turbo model.
- Our attack preserves or improves text quality (lower perplexity) while using much weaker reference models (e.g., using OPT-125M for OPT-30B).

Paper

Contact me