Can Watermarking Large Language Models Prevent Copyrighted Text Generation and Hide Training Data?

Michael-Andrei Panaitescu-Liess
mpanaite@umd.edu

Zora Che

Bang An

Yuancheng Xu

Pan Pathmanathan

Souradip Charkraborty

Sicheng Zhu

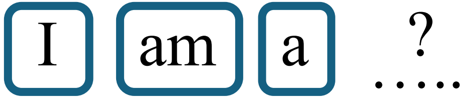
Tom Goldstein

Furong Huang

Paper

University of Maryland, College Park

How does LLM watermarking work?



I am a

?

……

Paper

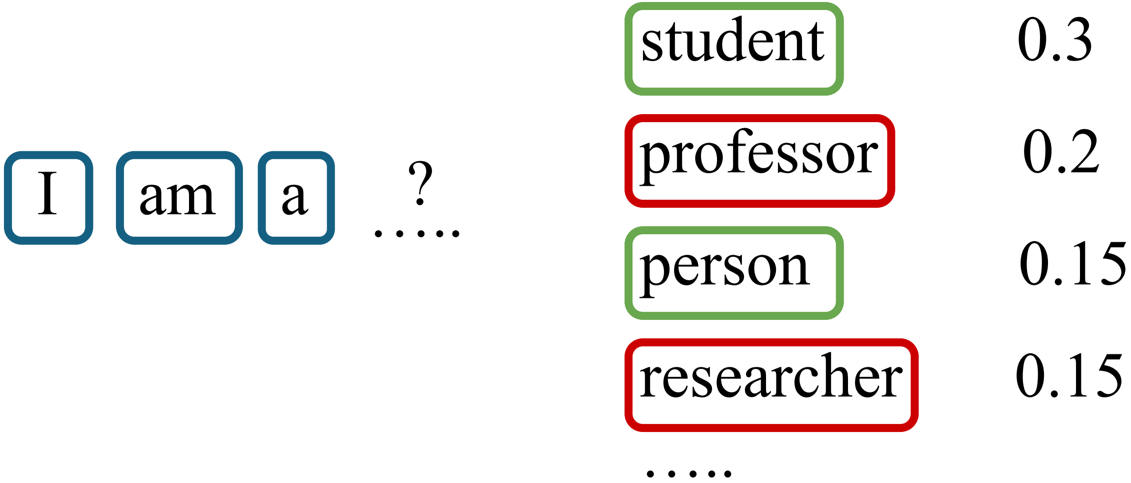
How does LLM watermarking work?

I am a ?
……

student	0.3
professor	0.2
person	0.15
researcher	0.15
…..	

Paper

How does LLM watermarking work?



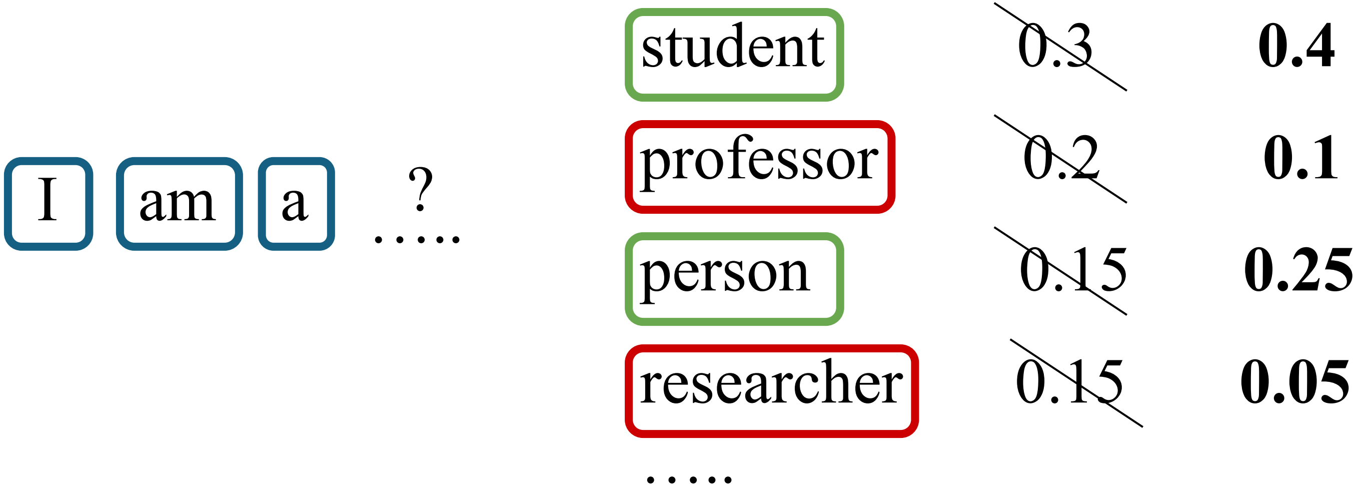
I am a ? ……

🟩	student	0.3
🟥	professor	0.2
🟩	person	0.15
🟥	researcher	0.15

……

Paper

How does LLM watermarking work?



I am a ?
.....

🟩 student	~~0.3~~	**0.4**
🟥 professor	~~0.2~~	**0.1**
🟩 person	~~0.15~~	**0.25**
🟥 researcher	~~0.15~~	**0.05**

.....

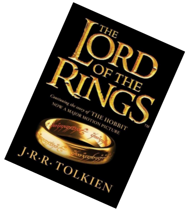
The model is biased towards green tokens.

Paper

[1] Kirchenbauer, et al. "A watermark for large language models."
[2] Zhao, et al. "Provable robust watermarking for ai-generated text."

✓ Watermarking reduces the generation of copyrighted text

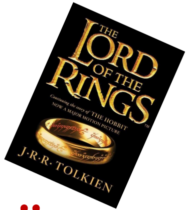
Prompt: "When Mr. Bilbo Baggins of Bag End announced that he would shortly"



Paper

✓ Watermarking reduces the generation of copyrighted text

Prompt: **"When Mr. Bilbo Baggins of Bag End announced that he would shortly"**



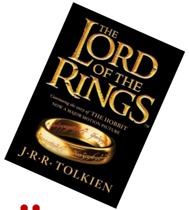
Completion for Llama-2-7b w/o Watermark (verbatim memorization):

"be celebrating his eleventy-first birthday with a party of special magnificence"

Paper

✓ Watermarking reduces the generation of copyrighted text

Prompt: "When Mr. Bilbo Baggins of Bag End announced that he would shortly"



Completion for Llama-2-7b w/o Watermark (verbatim memorization):

"be celebrating his eleventy-first birthday with a party of special magnificence"

Completion for Llama-2-7b w/ Watermark:

"become wealthy, and give a dinner to all his relatives and friends"

Paper

✓ Watermarking reduces the generation of copyrighted text

Similarity between the completion and the copyrighted text ©

	w/o watermark	**w/ watermark**
LLM trained on 50 copies of ©		

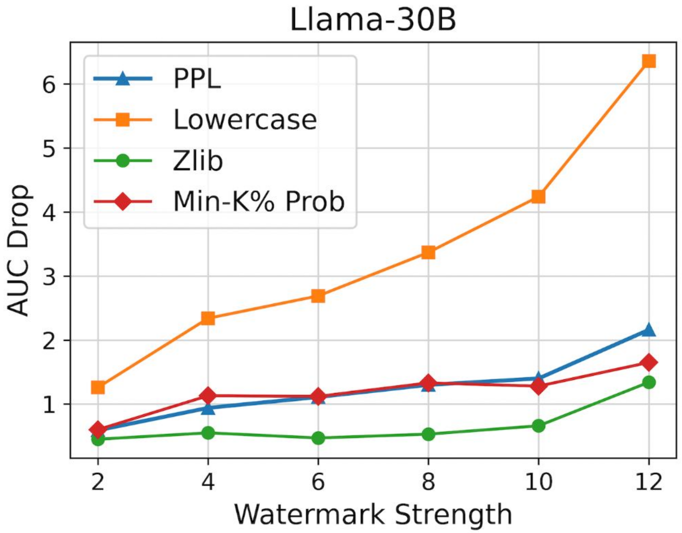
Paper

✓ Watermarking reduces the generation of copyrighted text

	Similarity between the completion and the copyrighted text ©	
	w/o watermark	**w/ watermark**
LLM trained on 50 copies of ©	**70%**	**24%**

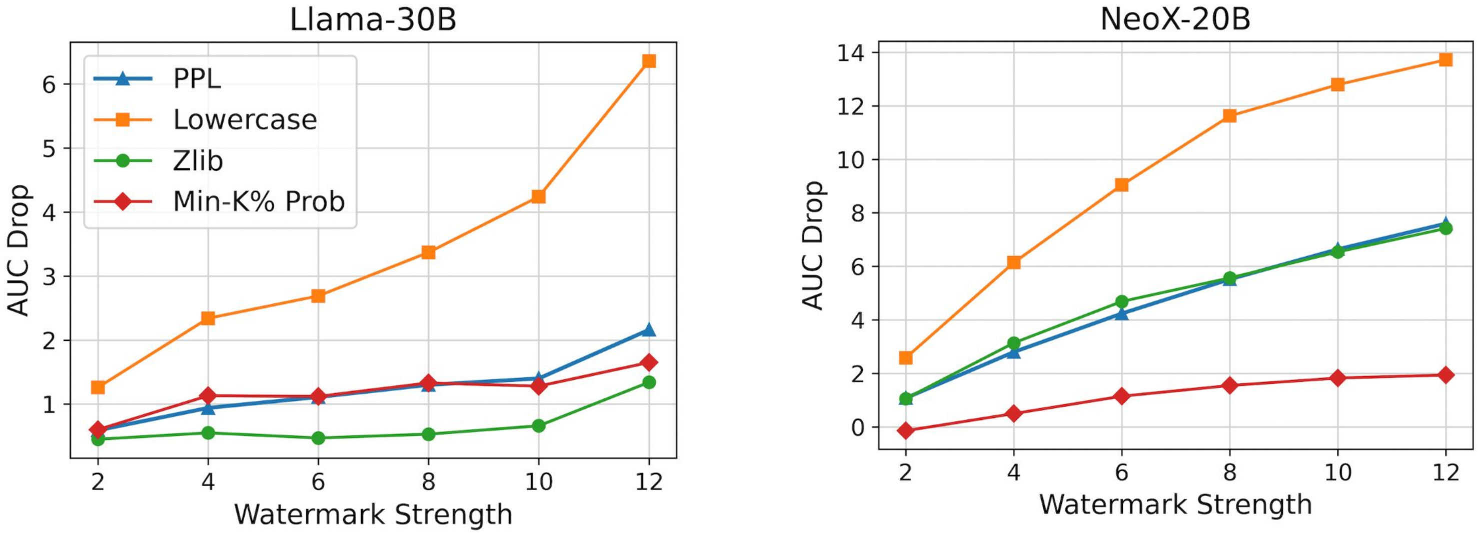
Paper

✗ Watermarking reduces the efficacy of training data detection methods



Paper

❌ Watermarking reduces the efficacy of training data detection methods



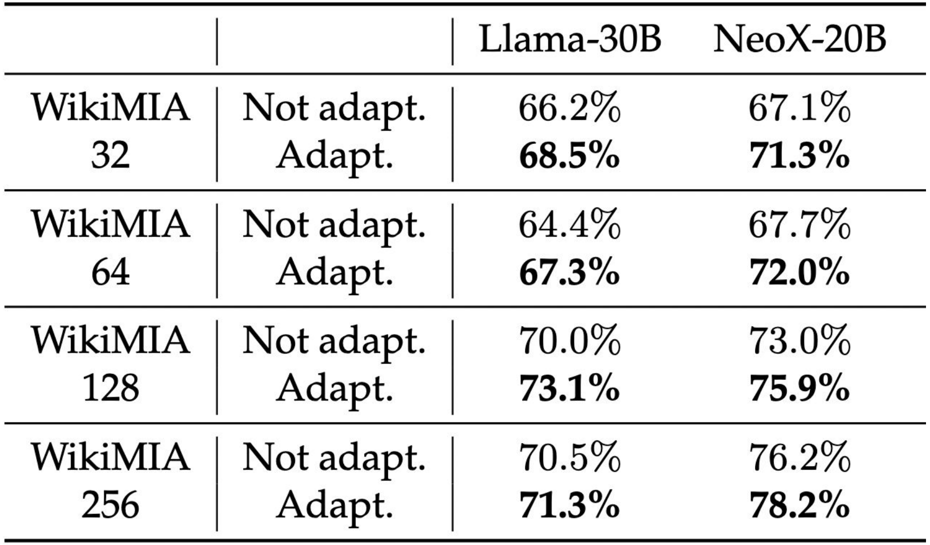
Paper

Is there anything we can do?

Paper

Is there anything we can do?

Adaptive methods improve the detection performance under watermarking



		Llama-30B	NeoX-20B
WikiMIA 32	Not adapt.	66.2%	67.1%
	Adapt.	**68.5%**	**71.3%**
WikiMIA 64	Not adapt.	64.4%	67.7%
	Adapt.	**67.3%**	**72.0%**
WikiMIA 128	Not adapt.	70.0%	73.0%
	Adapt.	**73.1%**	**75.9%**
WikiMIA 256	Not adapt.	70.5%	76.2%
	Adapt.	**71.3%**	**78.2%**

Paper

Watermarking can be a double-edged sword for copyright regulators since

- ***it promotes compliance during generation time,***
- ***but can make training time copyright violations harder to detect.***

Paper

Thank you for your attention!

Special thanks to the AdvML-Frontiers workshop organizers for their efforts!

Paper