谁用了我的模型？
—大模型水印前沿探索

G4-大模型伴生安全小组

任昱冰

2024.08.23

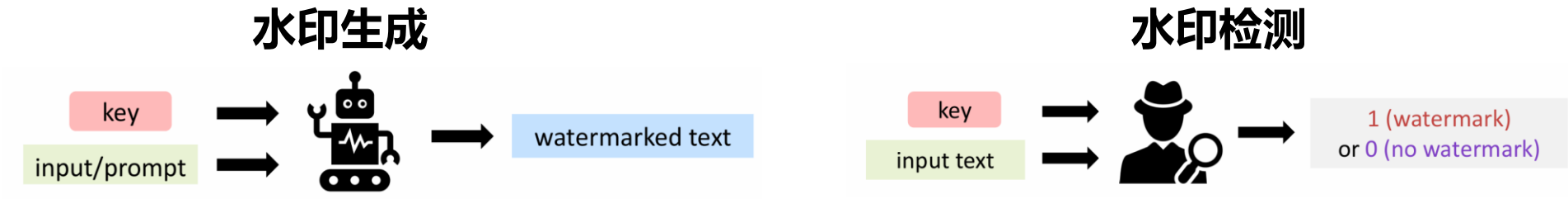
中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

提　纲

提　纲

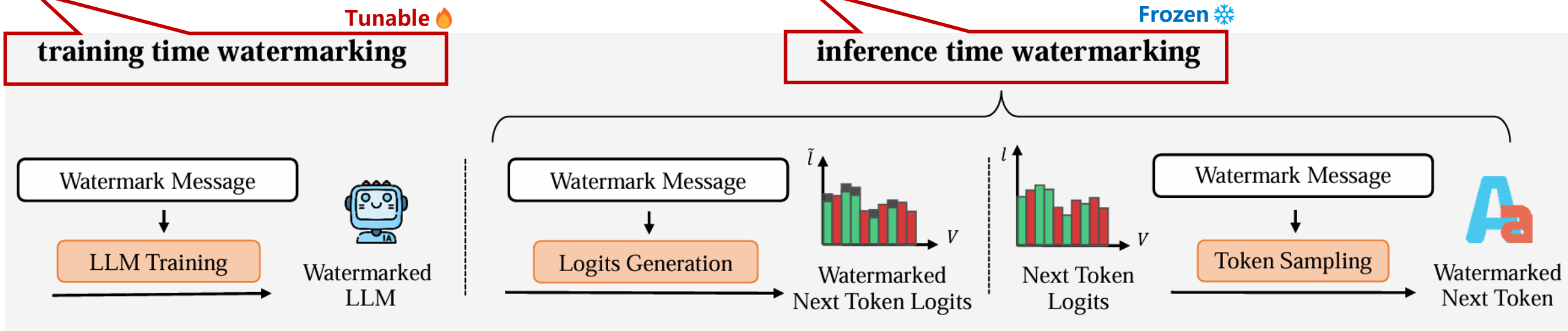
问题定义

■ 水印算法：



①让大模型学习隐式水印特征

②在模型输出文本中嵌入水印信息

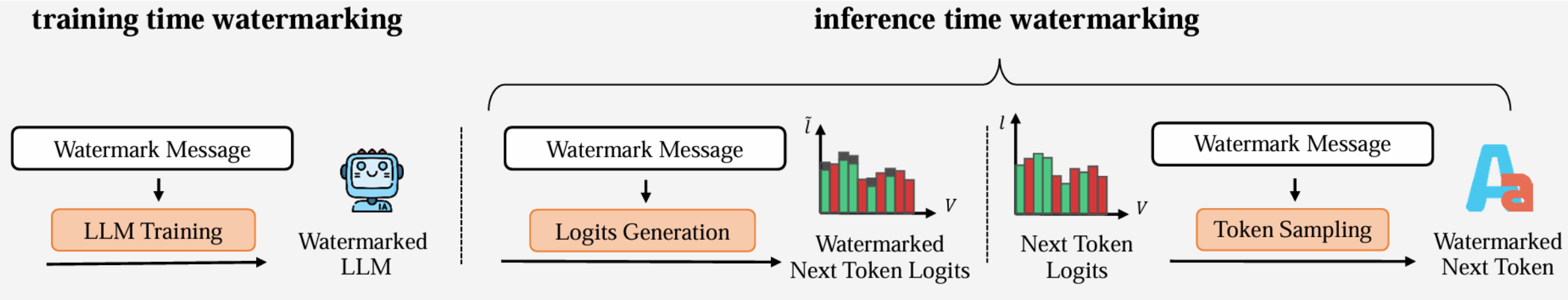


Aiwei Liu, et.al., A Survey of Text Watermarking in the Era of Large Language Models. 2024

提　纲

方法概况

- **Training** Time Watermarking
- **Inference** Time Watermarking
  - ① Watermarking During **Logits Generation**
  - ② Watermarking During **Token Sampling**



Aiwei Liu, et.al., A Survey of Text Watermarking in the Era of Large Language Models. 2024

Training Time Watermarking

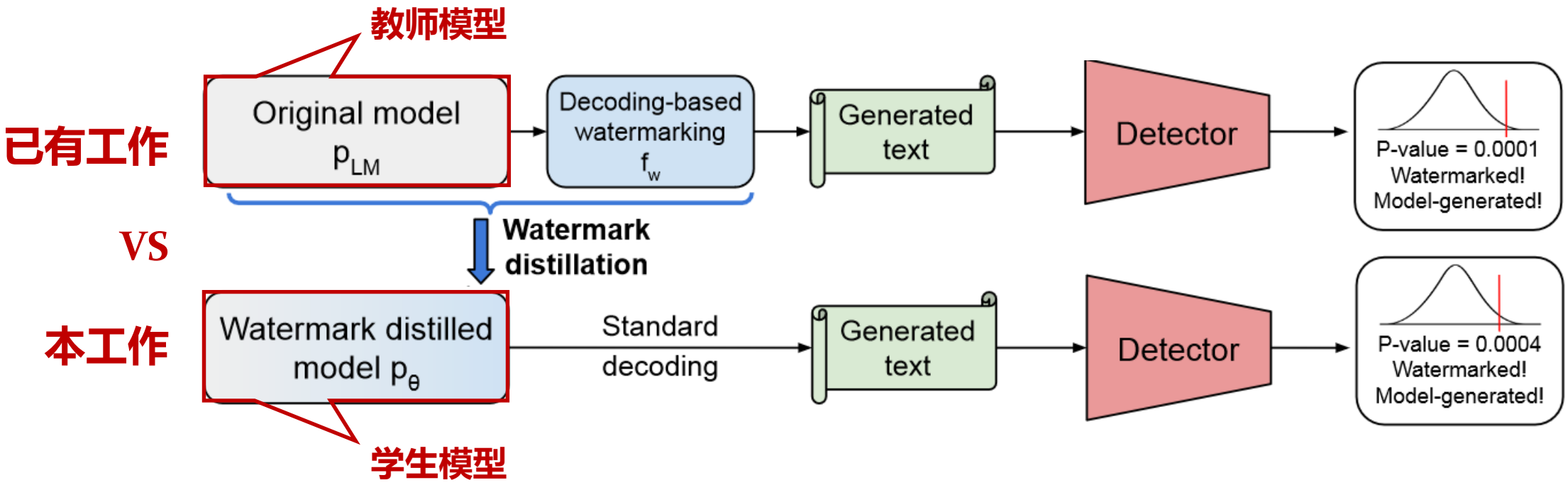
Published as a conference paper at ICLR 2024

ON THE LEARNABILITY OF WATERMARKS FOR LANGUAGE MODELS

Chenchen Gu, Xiang Lisa Li, Percy Liang, Tatsunori Hashimoto
Stanford University
{cygu, xlisali, thashim}@stanford.edu, pliang@cs.stanford.edu

Training Time Watermarking

- 探究语言模型水印的可学习性
- 将已有模型作为教师模型，通过**水印蒸馏**来训练新的学生模型，生成水印文本



Chenchen Gu, et.al., Stanford University, On the Learnability of Watermarks for Language Models (ICLR 2024)

Training Time Watermarking

1. Logit-based Watermark Distillation

训练目标：利用KL散度对齐教师模型 p_{LM} 和学生模型 p_θ 的分布

$$\mathcal{L}_{\text{logit}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \sum_{t=1}^{\text{len}(x)} D_{\text{KL}} \left(f_{\text{w}} \left(p_{\text{LM}} \left(\cdot \mid x_{<t} \right), x_{<t}, \xi \right) \parallel p_\theta \left(\cdot \mid x_{<t} \right) \right).$$

2. Sampling-based Watermark Distillation

训练目标：用教师模型 p_{LM} 生成部分水印文本并入训练集，利用交叉熵损失训练学生模型 p_θ

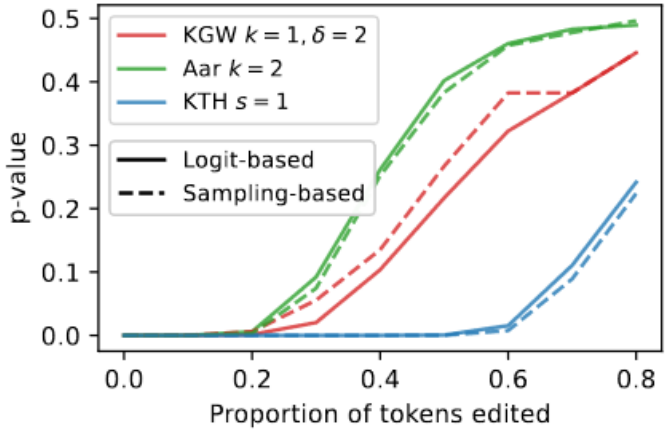
$$\mathcal{L}_{\text{sampling}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \sum_{t=1}^{\text{len}(x)} -\log p_\theta \left(x_t \mid x_{<t} \right).$$

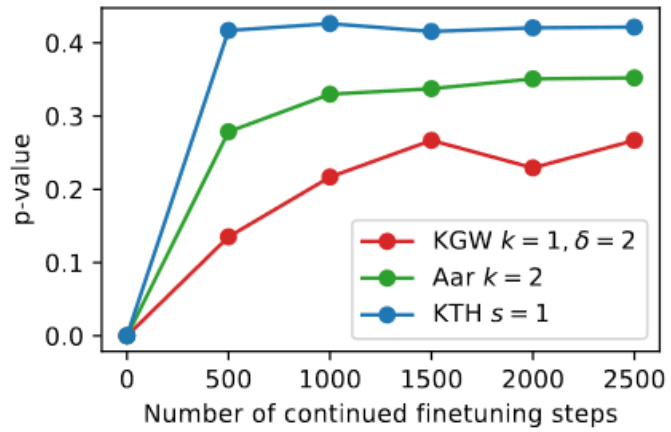
Chenchen Gu, et.al., Stanford University, On the Learnability of Watermarks for Language Models (ICLR 2024)

Training Time Watermarking

- **教师/学生模型：Llama2-7B**
- **数据集：C4**

Watermark		p-value (↓) (KTH test statistic (↓))			AUROC (↑)			Perplexity (↓)			seq-rep-3 (↓)		
		Decoding	Logit	Sampling	Decoding	Logit	Sampling	Decoding	Logit	Sampling	Decoding	Logit	Sampling
KGW	$k=0,\delta=2$	6e-16	2e-17	2e-15	1.00	1.00	1.00	17.5	17.3	20.3	0.05	0.05	0.05
	$k=1,\delta=2$	4e-18	7e-09	8e-07	1.00	1.00	1.00	16.5	17.6	19.2	0.04	0.03	0.04
	$k=2,\delta=2$	9e-18	1e-01	1e-01	1.00	0.80	0.74	16.8	17.7	19.8	0.03	0.02	0.03
	$k=0,\delta=1$	5e-04	3e-05	1e-03	0.98	0.99	0.98	13.0	12.9	15.7	0.03	0.03	0.03
	$k=1,\delta=1$	1e-05	7e-03	2e-02	0.99	0.91	0.87	12.7	13.1	14.9	0.03	0.03	0.03
Aar	$k=2$	1e-75	2e-12	3e-17	1.00	1.00	0.98	6.5	10.8	7.7	0.34	0.11	0.34
	$k=3$	5e-73	1e-01	6e-03	1.00	0.78	0.88	9.5	11.6	10.5	0.14	0.04	0.17
	$k=4$	4e-72	4e-01	3e-01	1.00	0.58	0.65	10.7	11.8	11.9	0.09	0.03	0.11
KTH	$s=1$	1e-04 (-593)	1e-04 (-565)	1e-04 (-561)	1.00	1.00	1.00	10.5	16.5	15.1	0.03	0.04	0.03
	$s=2$	1e-04 (-596)	1e-04 (-476)	1e-04 (-525)	1.00	0.99	0.99	10.7	16.3	13.4	0.03	0.04	0.03
	$s=4$	1e-04 (-594)	1e-03 (-438)	1e-04 (-487)	1.00	0.96	0.99	10.6	14.2	12.5	0.03	0.04	0.04
	$s=256$	1e-04 (-594)	8e-02 (-423)	1e-04 (-453)	1.00	0.85	0.97	10.8	11.3	11.3	0.03	0.04	0.04
Base student			5e-01			0.50			11.8			0.03	





Chenchen Gu, et.al., Stanford University, On the Learnability of Watermarks for Language Models (ICLR 2024)

Inference Time Watermarking - During Logits Generation

Published as a conference paper at ICLR 2024

UNBIASED WATERMARK FOR LARGE LANGUAGE MODELS

Zhengmian Hu[1], **Lichang Chen**[1], **Xidong Wu**[2], **Yihan Wu**[1], **Hongyang Zhang**[3], **Heng Huang**[1]
[1]Department of Computer Science, University of Maryland, College Park, MD 20742, USA
[2]Department of ECE, University of Pittsburgh, Pittsburgh, PA 15261, USA
[3]School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Zhengmian Hu, et.al., University of Maryland, Unbiased Watermark for Large Language Models (ICLR 2024 Spotlight)

Inference Time Watermarking - During Logits Generation

- **什么是无偏水印?**

Let P be the probability distribution of the original language model. A watermark function R with a random variable E (representing the watermark code) is unbiased if:

$$\mathbb{E}[R(P, E)] = P$$

where $\mathbb{E}$ is the expectation over E.

Key Point: An unbiased watermark function ensures that the expectation of the reweighted probabilities equals the original probabilities.

- **无偏水印方法:**

① 无偏赋权（Unbiased Reweight）② 独立水印码（Independent Watermark Codes）

Zhengmian Hu, et.al., University of Maryland, Unbiased Watermark for Large Language Models (ICLR 2024 Spotlight)

Inference Time Watermarking - During Logits Generation

- **无偏赋权：确保含水印分布的期望值与原始分布相匹配**

  ① δ-reweight：从原始分布中采样一个One-hot分布

  ② γ-reweight：随机重排词表，将概率分布范围减半，使剩余token的概率x2

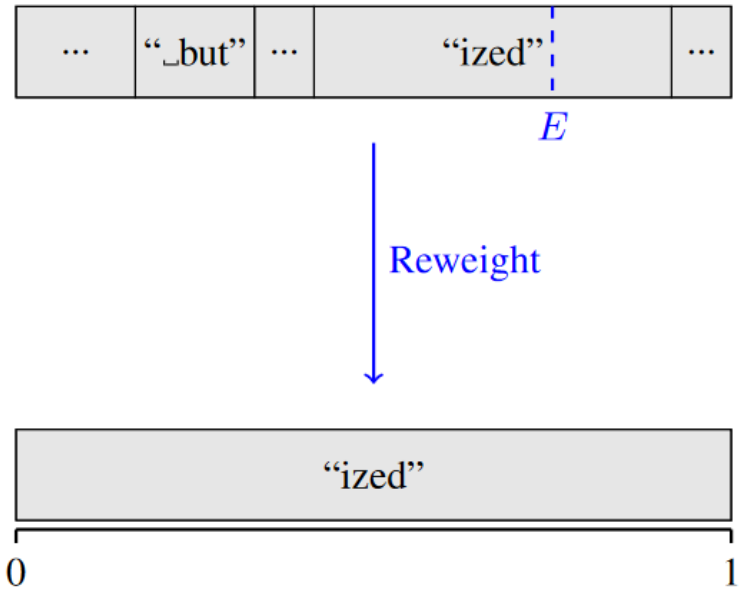


Figure 1: Illustration of δ-reweight.

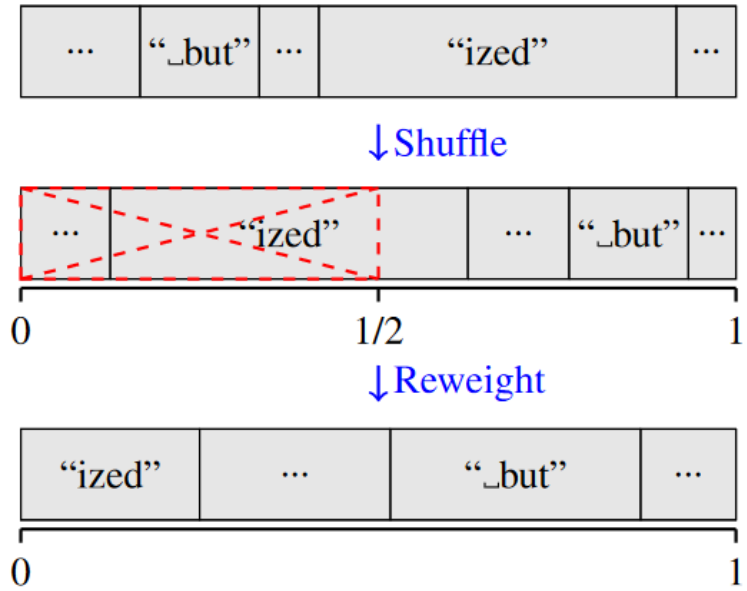
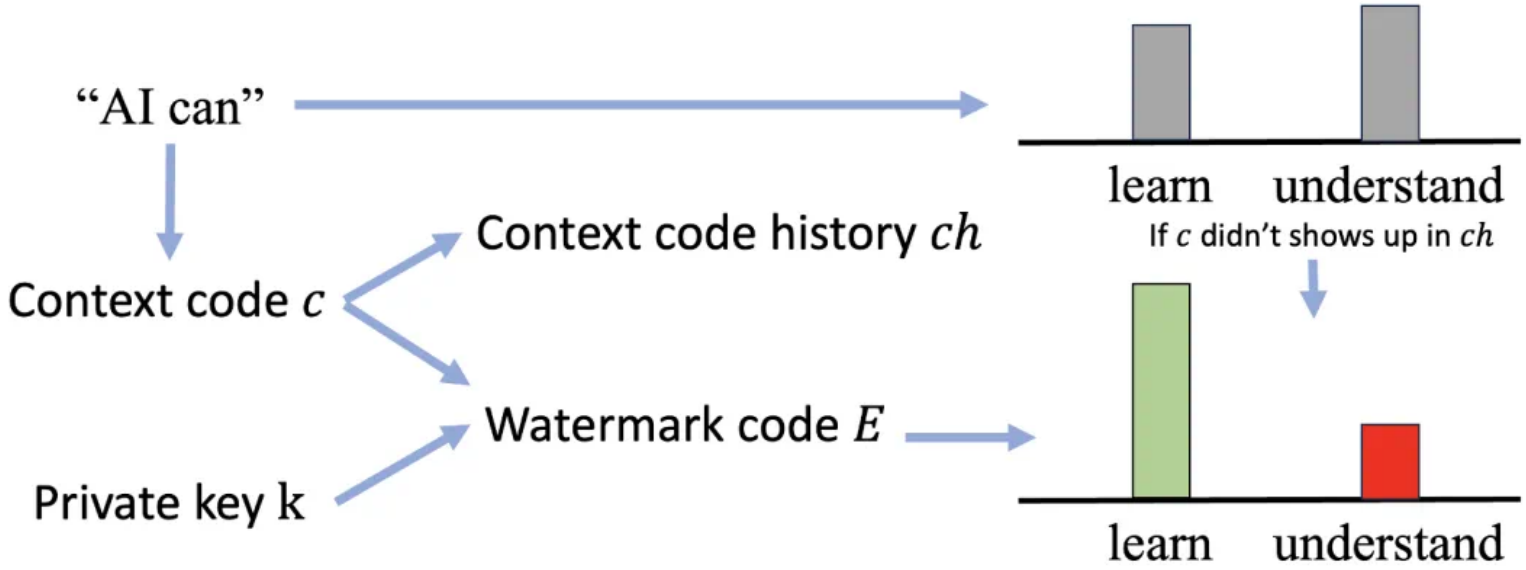


Figure 2: Illustration of γ-reweight.

Zhengmian Hu, et.al., University of Maryland, Unbiased Watermark for Large Language Models (ICLR 2024 Spotlight)

Inference Time Watermarking - During Logits Generation

- **独立水印码：保证整个序列的无偏性**



- **生成过程中，如果某个上下文码出现过，就跳过水印嵌入，直接使用原始的语言模型输出分布**

Zhengmian Hu, et.al., University of Maryland, Unbiased Watermark for Large Language Models (ICLR 2024 Spotlight)

Inference Time Watermarking - During Logits Generation

- **水印检测：**

① 基于似然的检测：对数似然比检验方法

比较给定文本在原始分布和水印分布下的似然，如果似然比超过一个阈值，则判定该文本含有水印

$$S = \log \frac{P_{M,w}(\boldsymbol{x}_{1:n}|\boldsymbol{a}_{1:m};k)}{P_M(\boldsymbol{x}_{1:n}|\boldsymbol{a}_{1:m})}$$

② 无需似然的检测

设计了一种新的打分函数

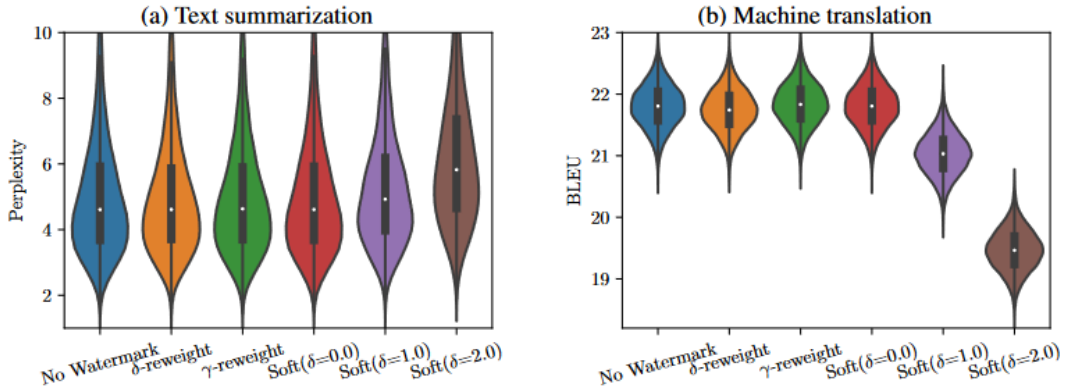
$$\max_{S_i} \min_{Q_i' \in \Delta_\Sigma, TV(Q_i', Q_i) \leq d} \langle Q_i', S_i \rangle, \quad s.t. \langle P_i, \exp(S_i) \rangle \leq 1.$$

Zhengmian Hu, et.al., University of Maryland, Unbiased Watermark for Large Language Models (ICLR 2024 Spotlight)

Inference Time Watermarking - During Logits Generation

- **实验：文本摘要，机器翻译**
- **嵌入水印后的文本质量与原始模型输出相当，在ROUGE、BLEU等指标上无显著差异**

	Text summarization			Machine translation	
	BERTScore ↑	ROUGE-1 ↑	Perplexity ↓	BERTScore ↑	BLEU ↑
No Watermark	32.70 ± 0.08	38.56 ± 0.09	5.024 ± 0.018	55.9 ± 0.3	21.8 ± 0.3
δ-reweight	32.71 ± 0.08	38.57 ± 0.09	5.022 ± 0.018	56.3 ± 0.3	21.7 ± 0.3
γ-reweight	32.69 ± 0.08	38.60 ± 0.09	5.019 ± 0.018	56.2 ± 0.3	21.8 ± 0.3
Soft(δ=0.0)	32.70 ± 0.08	38.56 ± 0.09	5.024 ± 0.018	55.9 ± 0.3	21.8 ± 0.3
Soft(δ=1.0)	32.35 ± 0.08	38.20 ± 0.09	5.313 ± 0.018	55.1 ± 0.3	21.0 ± 0.3
Soft(δ=2.0)	31.21 ± 0.08	37.17 ± 0.08	6.253 ± 0.022	53.8 ± 0.3	19.5 ± 0.3



Zhengmian Hu, et.al., University of Maryland, Unbiased Watermark for Large Language Models (ICLR 2024 Spotlight)

Inference Time Watermarking - During Logits Generation

Published as a conference paper at ICML 2024

Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models

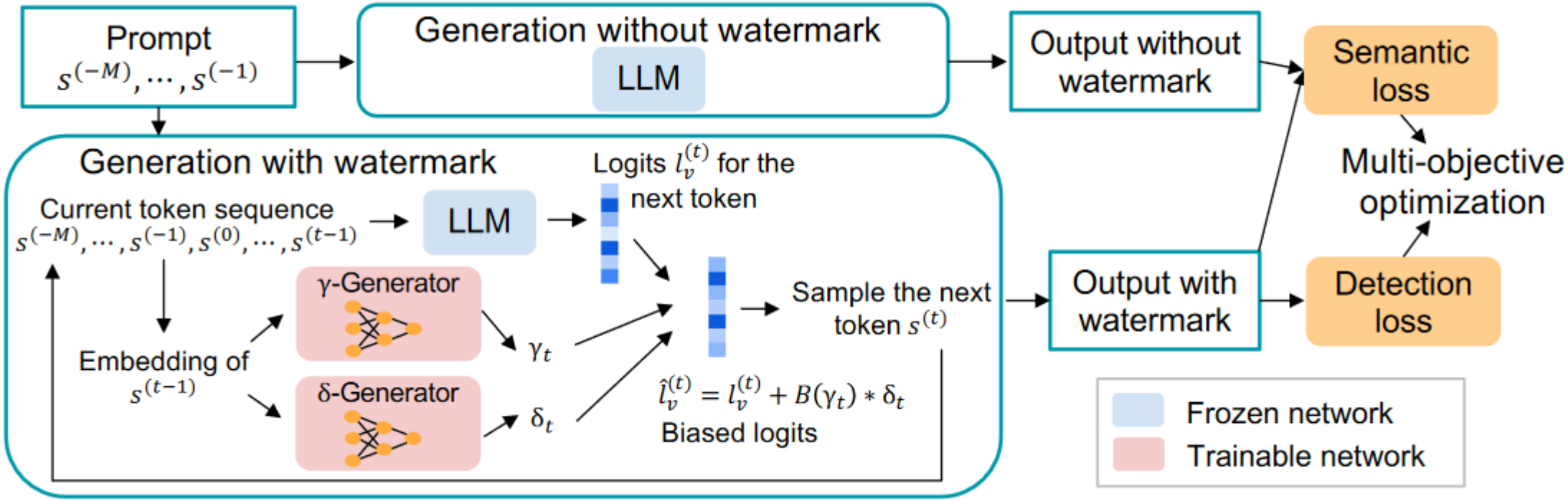
Mingjia Huo [*1] **Sai Ashish Somayajula** [*1] **Youwei Liang** [1] **Ruisi Zhang** [1] **Farinaz Koushanfar** [1] **Pengtao Xie** [1]

Department of Electrical and Computer Engineering, University of California.

Mingjia Huo, et.al., University of California, Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models (ICML 2024)

Inference Time Watermarking - During Logits Generation

- **改进KGW中绿表切分比例 γ 和logits扰动值 δ**



Mingjia Huo, et.al., University of California, Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models (ICML 2024)

Inference Time Watermarking - During Logits Generation

- **基座模型：OPT-1.3B, LLAMA2-7B/13B/70B**
- **数据集：C4 dataset**

Table 1: Comparison of EXP-edit and Our Method

Method	TPR @ 0%	TPR @ 1%	SimCSE
EXP-edit	0.922	0.996	0.655
EXP-edit (Top-k=50)	0.968	0.996	0.677
Ours (Top-k=50)	**1.000**	**1.000**	**0.713**

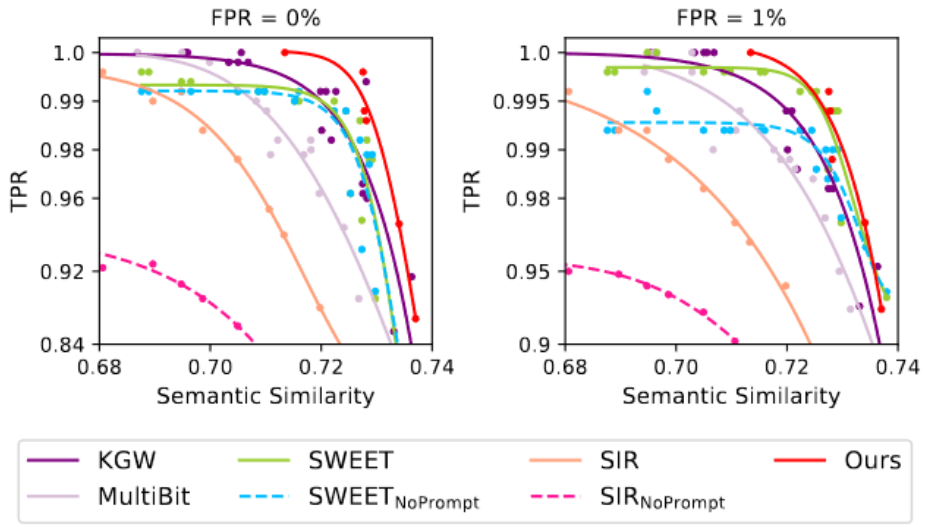


Figure 2: Comparison of the trade-off for semantic integrity and detectability of different methods applied to OPT-1.3B.

Mingjia Huo, et.al., University of California, Token-Specific Watermarking with Enhanced Detectability and Semantic Coherence for Large Language Models (ICML 2024)

Inference Time Watermarking - During Token Sampling

Robust Distortion-free Watermarks for Language Models

Rohith Kuditipudi John Thickstun Tatsunori Hashimoto Percy Liang

Department of Computer Science
Stanford University

July 2023

Rohith Kuditipudi, et.al., Stanford University, Robust Distortion-free Watermarks for Language Models (TMLR 2024)

Inference Time Watermarking - During Token Sampling

- **水印注入：用伪随机数生成器生成一系列伪随机数，以指导每个词元的采样过程**
- **水印检测：用编辑距离评估文本词元与伪随机数之间的对齐情况**

Algorithm 4: Randomized watermarked text generation (`shift-generate`)

Input : watermark key sequence $\xi \in \Xi^n$
Params: generation length m, language model p, decoder Γ
Output: string $y \in \mathcal{V}^m$

$\tau \sim \text{Unif}([n])$, $\xi' \leftarrow \{\xi_{(i+\tau)\%n}\}_{i=1}^m$
return `generate`$(\xi'; m, p, \Gamma)$

Algorithm 2: Watermarked text detection (`detect`)

Input : string $y \in \mathcal{V}^*$, watermark key sequence $\xi \in \Xi^n$
Params: test statistic ϕ; watermark key sequence distribution $\nu \in \Delta(\Xi^n)$; resample size T
Output: p-value $\widehat{p} \in [0, 1]$

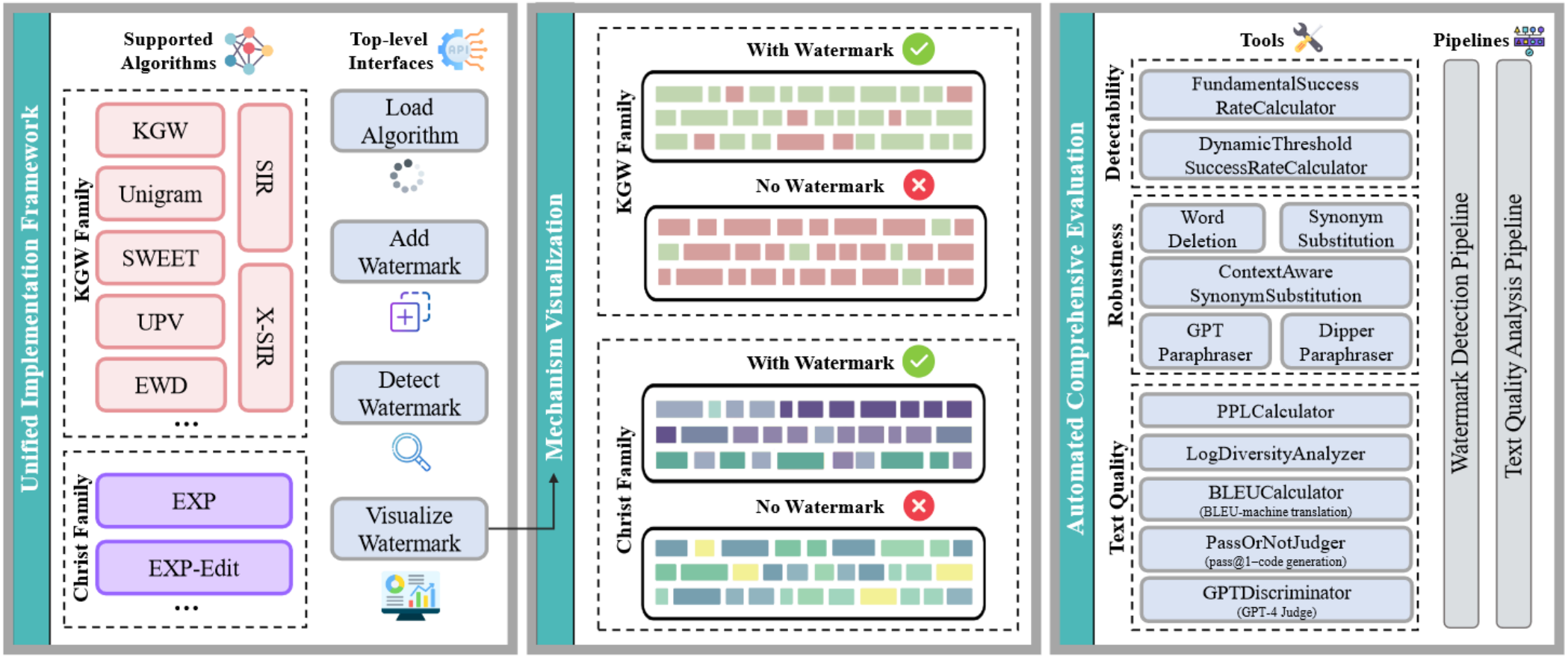
for $t \in 1, \ldots, T$ **do**
  $\xi^{(t)} \sim \nu$
  $\phi_t \leftarrow \phi(y, \xi^{(t)})$
$\widehat{p} \leftarrow \frac{1}{T+1}\left(1 + \sum_{t=1}^{T} \mathbf{1}\{\phi_t \leq \phi(y, \xi)\}\right)$
return $\widehat{p}$

Rohith Kuditipudi, et.al., Stanford University, Robust Distortion-free Watermarks for Language Models (TMLR 2024)

提　纲

新视角

- **大模型水印开源工具：** https://github.com/THU- BPM/MarkLLM



Leyi Pan, et.al., 清华, MarkLLM: An Open-Source Toolkit for LLM Watermarking

新视角

- **大模型水印窃取：**

  查询水印模型API得到水印文本，逆向工程得到水印规则

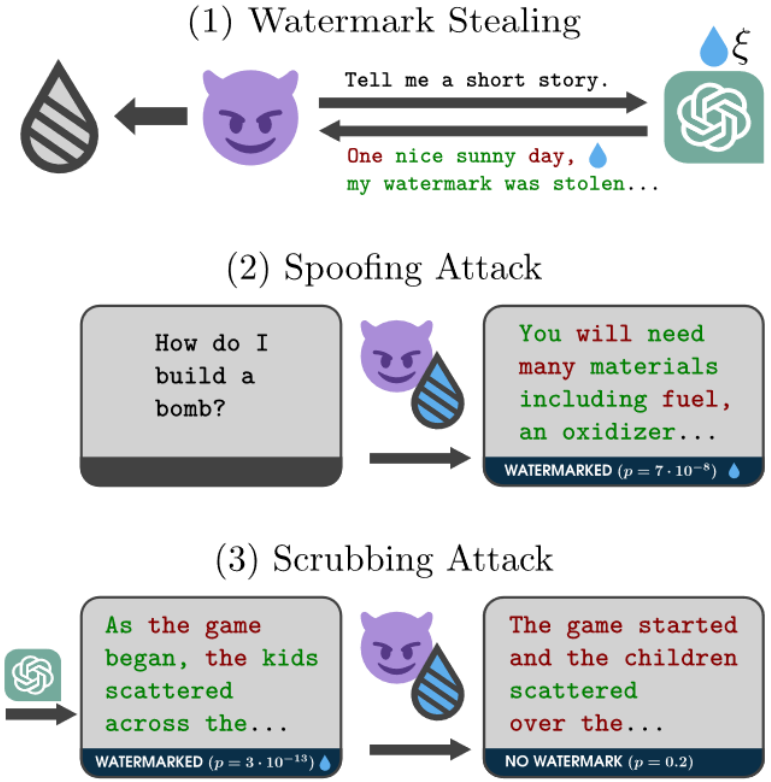
  目标：KGW

- **欺骗攻击：**

  攻击者在不知道密钥的情况下，仍能生成能被检测为带水印的文本

- **清洗攻击：**

  攻击者从水印文本中去除水印，产生一个被检测为无水印的有效释义



Nikola Jovanovi´ c, et.al., ETH Zurich, Watermark Stealing in Large Language Models (ICML 2024)

提　纲

总结展望

01

质量与版权的平衡

- 不会显著改变内容质量或可读性
- 水印必须隐蔽嵌入到生成内容中

02

应对复杂攻击场景

- 抵抗多种形式的攻击，如增删改
- 不轻易破坏水印完整性和可识别性

03

多比特水印内容

- 每个数字内容携带丰富的元数据
- 嵌入生成模型、版权归属等信息

04

适应多种生成任务

- 应用于翻译、对话等多种生成任务
- 适配不同应用场景和多种文本风格

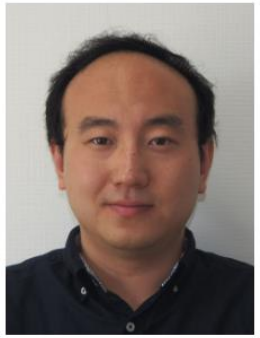
国内团队

姓名： 闻立杰
职称： 长聘副教授
职务： 党委学生工作组组长
邮件： wenlj@tsinghua.edu.cn
房间： 清华大学东配楼11-405
电话： 010-62783610
传真： 010-62781776

研究领域： 流程挖掘、大数据处理与分析、自然语言处理

个人简介： 河北唐山人，清华大学软件学院长聘副教授，博士生导师。在ACL、AAAI、SIGIR、SIGKDD、ASE、EMNLP、COLING、NAACL、BPM、CAiSE、SDM、CIKM、IEEE TSC、DMKD、DKE等发表论文170余篇，谷歌学术引用4300余次。主持国家重点研发计划课题2项、主持国家自然科学基金2项、参与国家NSFC/973/863计划子课题十余项、国家核高基重大专项课题1项。获国家发明专利、软件著作权十余项，获业务流程管理领域国际顶级会议BPM 2015最佳学生论文奖（亚洲首次，中国唯一）、CBPM 2017/2018/2020/2021最佳(学生)论文奖。流程挖掘论文已被收入国际教材和学术专著，流程管理领域译著3部。现任国际会议ACL、AAAI、EMNLP、CAiSE、ICSOC、BPM程序委员会委员，中国业务流程管理大会CBPM指导委员会执行主席、IEEE流程挖掘工作组XES标准化小组委员（中国唯一），曾任BPM程序委员会资深委员。主导研发交互式大数据处理与分析平台FloK和流程挖掘工具THUMiner，研究成果已在中国移动、华为、中国气象局、天远科技、中车四方所、辽宁瑞华等企业获得成功应用。

清华闻立杰



Rui Wang

上交王瑞

Associate Professor
Department of Computer Science and Engineering
Shanghai Jiao Tong University

Email: wangrui12 (as you know) sjtu.edu.cn

教授，博导，上海市数据科学重点实验室主任
地址：中国上海杨浦区淞沪路2005号江湾校区2号交叉学科楼
E-mail：shawyh@fudan.edu.cn
Tel: +86-021-51355548

复旦肖仰华

个人简介

复旦大学计算机科学技术学院教授、博导、上海市数据科学重点实验室主任。2009年获得复旦大学博士学位后留校任教，先后任讲师、副教授、教授（2017年）。

万小军

职称：教授
所在院系：王选计算机研究所
研究领域：自然语言处理、文本挖掘
办公电话：86-10-82529548
电子邮件：wanxiaojun@pku.edu.cn
个人主页：
https://wanxiaojun.github.io

北大王选万小军

北京大学王选计算机研究所研究员、博士生导师，语言计算与互联网挖掘研究室负责人，在北京大学获得学士、硕士与博士学位。研究方向为自然语言处理与文本挖掘，研究兴趣包括自动文摘、文本生成、情感分析、语义分析、多模态与多语言NLP等。曾担任计算语言学顶级国际期刊Computational Linguistics编委、国际会议EMNLP-IJCNLP 2019程序委员会主席，现任CCF-NLP专委会秘书长、TACL执行编辑、NLE编委、JCST编委，10多次担任相关领域重要国际会议领域主席，包括ACL、NAACL、EMNLP、EACL，AACL等。荣获ACL2017杰出论文奖、IJCAI 2018杰出论文奖、2017年吴文俊人工智能技术发明奖、CCF NLPCC青年新锐奖等奖励。研制推出了多款AI写作机器人，如小明、小南、小柯等，应用于多家媒体单位。

国内团队

熊德意

天大熊得意

姓名：熊德意

职称：教授

所在系别：计算机科学与技术学院

导师类型：博士生导师

电子邮件：dyxiong@tju.edu.cn

研究领域：自然语言处理

研究方向：机器翻译、对话、自然语言生成、机器阅读理解和问答、信息抽取、知识图谱

个人主页：https://dyxiong.github.io/

Juanzi Li

清华李涓子

Home Teaching Book

Professor

The principal of Knowledge Engineering Group

Research Interests: Knowledge Engineering and Semantic Web, Text and Social Network Mining

PhD Dissertation: (1996.9-2000.1) Chinese Word Sense Disambiguation, Supervisor: Prof. Huang Changning

Postdoctoral Thesis: (2000.1-2001.12) Chinese Structural Language Understanding Model, Coordinator: Prof. Wang Zuoying

Phone: (010)62781461 **Fax:** (010)62789831

Email: lijuanzi@tsinghua.edu.cn

Address: Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

邱锡鹏

教授，博士生导师，复旦大学计算机科学技术学院

复旦邱锡鹏

个人简介

于复旦大学获得理学学士和博士学位。研究方向为自然语言处理、大语言模型，发表C
型MOSS [GitHub]、开源自然语言处理工具FudanNLP [GitHub] [Google Code]、Fast
的广泛使用。指导学生多次获得中国人工智能学会优博、中国中文信息学会优博、微软
等。

(博士后招聘) (研究生招生说明) (科研工程助理招聘)

研究方向

围绕下一代大语言模型开展研究，包括大模型预训练、微调、对齐、轻量化、多模态融合

具体工作可参考：研究方向与代表性论文

Fandong Meng (孟凡东)

微信孟凡冬

Personal Information

He is a principal researcher and team leader at WeChat, Tencent Inc, China.

He got his Ph.D. degree at Institute of Computing Technology, Chinese Academy of Sciences.

His research interests include machine translation, natural language processing and large language modeling.

E-mail: fandongmeng [at] tencent [dot] com

国内团队

张卫明

中科大张卫明

中国科学技术大学 教授

1999年毕业于解放军信息工程大学，并于2005年在该校获得博士学位。现任中国科学技术大学教授，博导，网络空间安全学院副院长，图象图形学学会多媒体取证与安全专家委员会副秘书长。主要研究兴趣包括信息隐藏、数据隐私保护和人工智能安全。已在TIT、TIFS、TIP、TVCG、TCOMM、TDSC、CVPR、ICCV、AAAI、INFOCOM、ACM MM等期刊和会议发表论文200多篇。 获安徽省自然科学奖一等奖1项、安徽省教学成果特等奖1项。 主持国家重点研发项目课题、国家自然科学基金重点、面上、国家863等项目20余项。

目前他的H因子是30 (谷歌学术，2020年8月数据)，谷歌学术个人主页是 点击这里，ResearchGate主页是 点击这里。

谢谢，敬请批评指正！

Q&A

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS