Robust Utility-Preserving Text Anonymization Based on Large Language Models

Tianyu Yang[1], Xiaodan Zhu[1,2], Iryna Gurevych[1]

[1]Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science and Hessian Center for AI (hessian.AI), Technical University of Darmstadt, Germany

[2]Department of Electrical and Computer Engineering, Ingenuity Labs Research Institute, Queen's University, Canada

Abstract

Anonymizing text that contains sensitive information is crucial for a wide range of applications. Existing techniques face the emerging challenges of the re-identification ability of large language models (LLMs), which have shown advanced capability in memorizing detailed information and reasoning over dispersed pieces of patterns to draw conclusions. When defending against LLM-based re-identification, anonymization could jeopardize the utility of the resulting anonymized data in downstream tasks. In general, the interaction between anonymization and data utility requires a deeper understanding within the context of LLMs. In this paper, we propose a framework composed of three key LLM-based components: **a privacy evaluator**, **a utility evaluator**, and **an optimization component**, which work collaboratively to perform anonymization. Extensive experiments demonstrate that the proposed model outperforms existing baselines, showing robustness in reducing the risk of re-identification while preserving greater data utility in downstream tasks. We provide detailed studies on these core modules. To consider large-scale and real-time applications, we investigate the distillation of the anonymization capabilities into lightweight models.

Motivations

➢ Fundamental Challenge: Privacy-utility Tradeoff

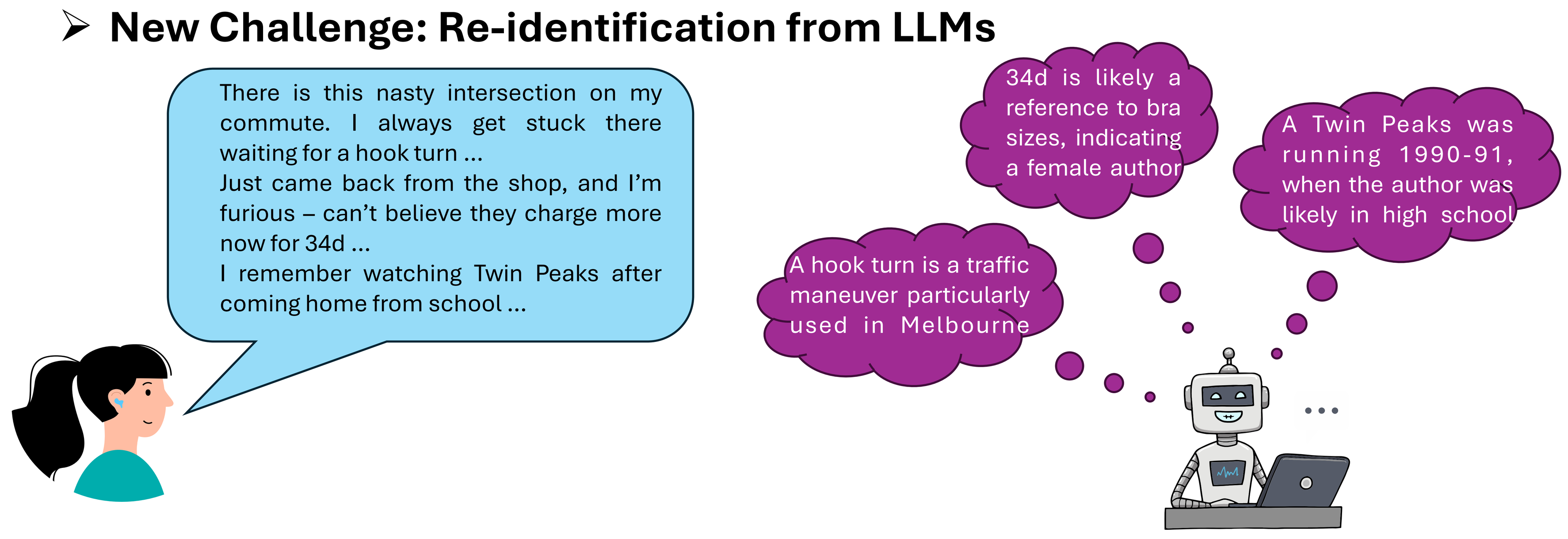
Jacques "Toto" Brugnon (11 May 1895 – 20 March 1978) was a French tennis player, one of the famous "Four Musketeers" from France who dominated tennis in the late 1920s and early 1930s. He was born in Paris and died in Paris. He was primarily a doubles specialist who won 10 Grand Slam doubles titles in the French, American, Australian and British championships ...

Original Text

Dsadd ddsad ds tennisdsdsa sdad dsdd pecializing dsadd dsadsd asdgghh fdsaf fds gfdsg several top fdsf titlfsdf es dsaf sad ds tennisdsdsa sdad dsdd pecializing dsadd dsadsd asdgghh fdsaf fds gfdsg several top fdsf titlfsdf es dsaf sad ds tennisdsdsa sdad dsdd pecializing dsadd dsadsd asdgghh fdsaf fds gfdsg several top fdsf titlfsdf es dsaf ...

Perfectly Anonymized Text ?

➢ New Challenge: Re-identification from LLMs



Methods

➢ Modeling Text Anonymization with Multi-objective Optimization

- ➢ We propose a novel framework for text anonymization that is built on the powerful ability of LLMs, consisting of a privacy evaluator, a utility evaluator, and an optimizer component. They work in tandem and show superior performance over the existing models.

➢ DPO-based Knowledge Distillation

- ➢ To provide a practical model for real-time environments, we investigate the distillation of anonymization capabilities into smaller models.

➢ New Benchmark

- ➢ We create a new dataset using the celebrity biographies from Dbpedia with occupation labels, serving as a practical benchmark for evaluating the impact of anonymization methods on downstream tasks. Anonymization results from LLMs are also included to aid future text anonymization research.

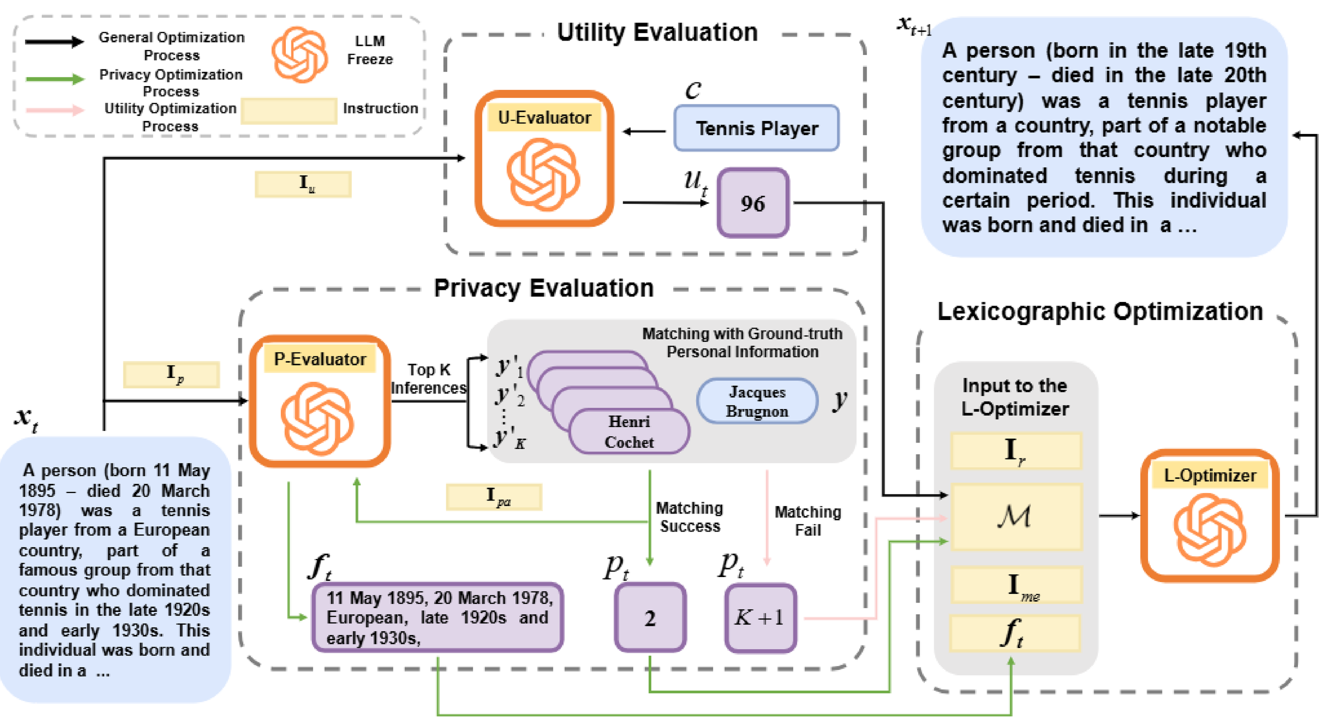


Figure 2: An overview of the proposed RUPTA framework. $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ denote the input and output text in one iteration; y denotes the ground-truth personal information; and $\boldsymbol{f}_t$, $[y_t^i]_1^K$ and p_t are the inference feedback, inferred personal information from P-Evaluator and the value of the privacy objective. The ground-truth downstream task label is denoted as c, while u_t is the value of the utility objective. $\mathcal{M}$ denotes the history optimization results. $\mathbf{I}_u, \mathbf{I}_p, \mathbf{I}_r, \mathbf{I}_{me}$ and $\mathbf{I}_{pa}$ are the prompts used for each component of the method.

Experiments

➢ Evaluation Results on the DB-bio Dataset

	Method	Disclosure Risk		Utility Preservation				
		SR⇓	CS⇓	Precision⇑	Recall⇑	F1⇑	Accuracy⇑	Loss⇓
DB-bio	**Original**	100.00	98.45	99.58	99.68	99.61	99.58	0.0422
	Azure (Aahill, 2023)	78.24	80.87	91.63	95.04	92.39	92.47	0.3202
	DEID-GPT (Liu et al., 2023)	77.10	79.47	90.82	94.37	92.56	91.22	0.3103
	SD (Dou et al., 2023)	73.21	73.63	92.27	93.11	92.69	92.96	0.2719
	IncogniText (Frikha et al., 2024)	58.06	56.28	85.68	89.03	87.32	88.28	0.4842
	AF (Staab et al., 2024b)	<u>52.91</u>	**50.84**	91.20	94.26	91.75	92.02	0.4048
	RUPTA (Mixtral 8×22b)	67.78	67.15	**96.18**	**97.13**	**96.30**	**96.23**	<u>0.2167</u>
	RUPTA (Llama-3-70b)	64.02	63.23	95.34	96.23	95.55	95.82	0.2224
	RUPTA (GPT-3.5)	68.51	69.16	95.40	96.02	95.70	95.49	0.2188
	RUPTA (GPT-4)	**52.67**	<u>53.11</u>†	<u>95.58</u>†	<u>96.26</u>†	<u>95.91</u>†	<u>96.02</u>†	**0.1618**†

Table 1: The main experiment results on the test set of the DB-bio dataset. The top and second performances are highlighted with bold font and underlined, respectively. Results of RUPTA (GPT-4) denoted by † are significantly better than that of the AF method under the one-tailed paired t-test ($p < 0.05$).

➢ Evaluation Results on the PersonaReddit Dataset

	Method	Disclosure Risk		Utility Preservation				
		SR⇓	CS⇓	Precision⇑	Recall⇑	F1⇑	Accuracy⇑	Loss⇓
Personal Reddit	**Original**	49.76	81.89	55.13	63.51	55.80	58.45	1.5695
	Azure (Aahill, 2023)	45.89	81.07	<u>54.04</u>	**58.49**	<u>54.17</u>	**57.00**	**1.7340**
	DEID-GPT (Liu et al., 2023)	43.12	72.81	53.98	58.21	54.06	56.31	1.9314
	SD (Dou et al., 2023)	44.05	75.17	**54.11**	<u>58.43</u>	**54.21**	<u>56.93</u>	<u>1.7501</u>
	IncogniText (Frikha et al., 2024)	37.55	60.02	10.19	11.06	10.61	13.47	4.4766
	AF (Staab et al., 2024b)	<u>35.40</u>	<u>57.76</u>	16.64	22.32	16.68	21.26	3.3380
	RUPTA (Mixtral 8×22b)	**35.27**	65.56	37.37	47.82	37.67	43.48	2.2836
	RUPTA (Llama-3-70b)	39.61	61.63	32.96	44.57	32.82	38.65	2.3131
	RUPTA (GPT-3.5)	34.30	61.50	32.04	40.44	31.97	36.23	2.4477
	RUPTA (GPT-4)	35.75	**55.04**	30.34†	39.14†	30.09†	35.75†	2.5391†

Table 2: Experimental results on the test set of the PersonalReddit dataset. The top and second performances are highlighted with bold font and underlined, respectively. Results of RUPTA (GPT-4) denoted by † are significantly better than that of the AF method under the one-tailed paired t-test ($p < 0.05$).

➢ Anonymization Process

t	1	2	3	4
u_t	43.14	42.31	43.30	44.08

Table 3: Average u_t score during the anonymization process on the PersonalReddit dataset.

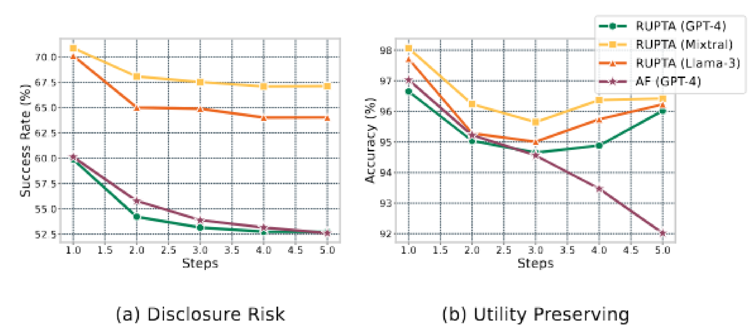


Figure 3: Evaluation results of the anonymized text at each iteration during the anonymization process using the AF and RUPTA methods with GPT-4, Llama-3-70b (Llama-3), and Mixtral 8 × 22b (Mixtral) as optimizers on the test set of the DB-bio dataset.

➢ Knowledge Distillation Experiments

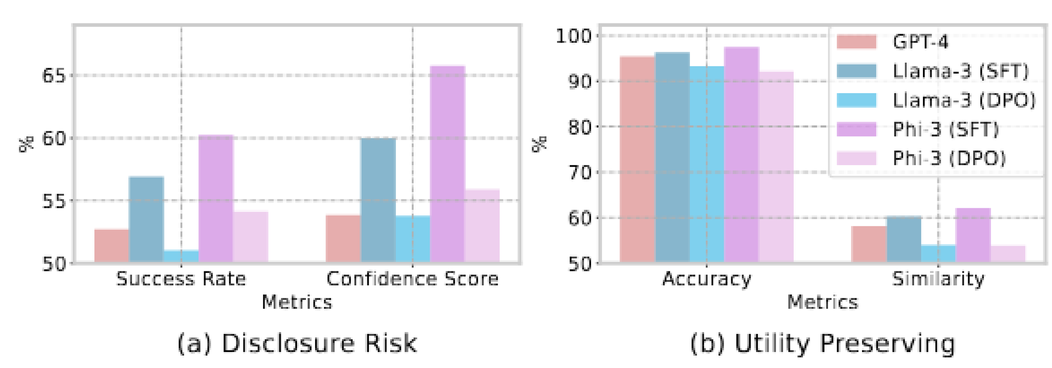


Figure 5: Results of the knowledge distillation experiment using Llama-3-8b (Llama-3) and Phi-3 Mini (Phi-3) as the student model, respectively.

Code **Contact**