NEURAL INFORMATION PROCESSING SYSTEMS

Self-Refining Language Model Anonymizers via Adversarial Distillation

Kyuyoung Kim[1*], Hyunjun Jeon[1*], Jinwoo Shin[1]

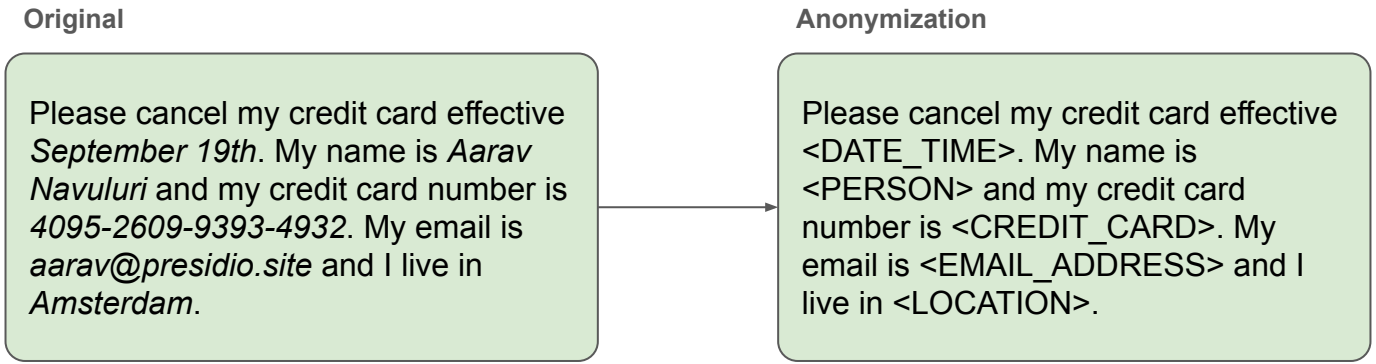
[1] KAIST

KAIST

Background

Text anonymization aims to remove or obscure PII while preserving the overall semantics and utility.

Traditional methods primarily remove *explicit* identifiers (e.g., names, SSNs, locations) through named entity recognition or pattern matching.



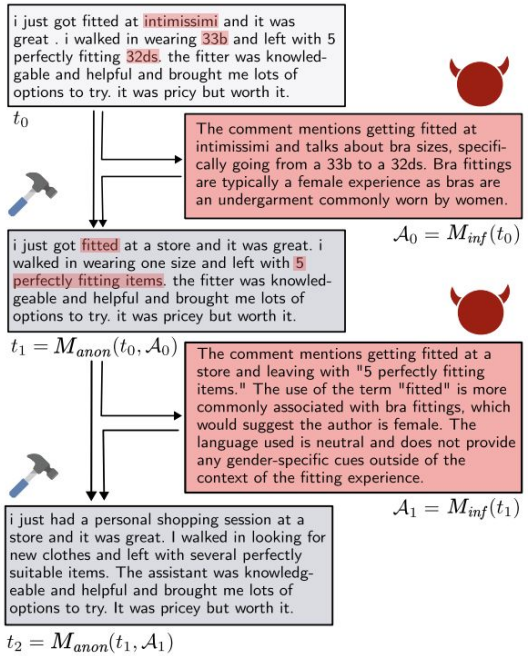
Background

Staab et al. propose leveraging **LLMs** for iteratively anonymizing such contextually embedded PII [1].

- $\mathbf{M}_{\text{anon}}$ generates an anonymization of an input text.
- $\mathbf{M}_{\text{inf}}$ attempts to infer PII from the anonymization.
- Based on the inference, $\mathbf{M}_{\text{anon}}$ further processes the text.

Using **GPT-4** for anonymization and inference proves significantly more effective than traditional methods.

However, using proprietary models is costly and poses privacy risks by exposing sensitive data to external systems.

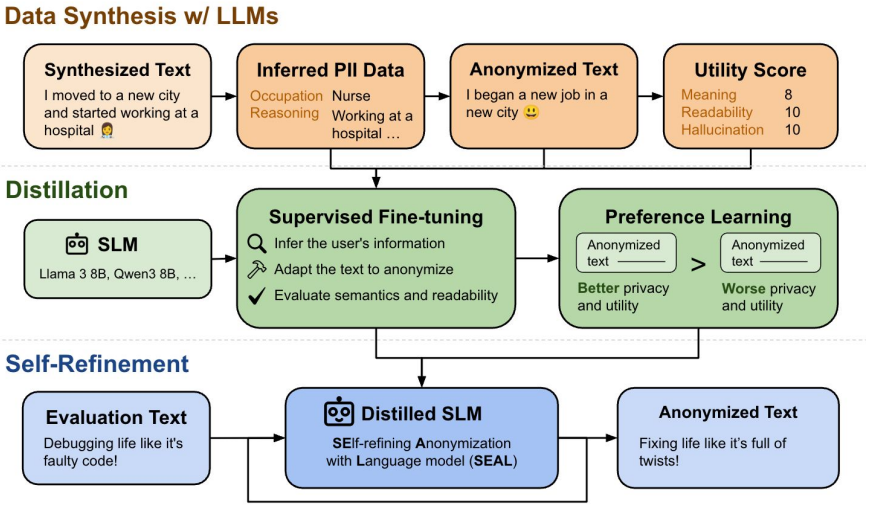


[1] [Staab et al., 24] Language Models are Advanced Anonymizers.

Self-Refining Anonymization

Question. Can we leverage **adversarial anonymization** (AA) with LLMs to collect data for **distilling** anonymization capabilities into SLMs?

- Specifically, we want to train SLMs that can both 1) anonymize and 2) evaluate their own outputs for privacy and utility.
- This enables anonymization at inference without relying on external models.

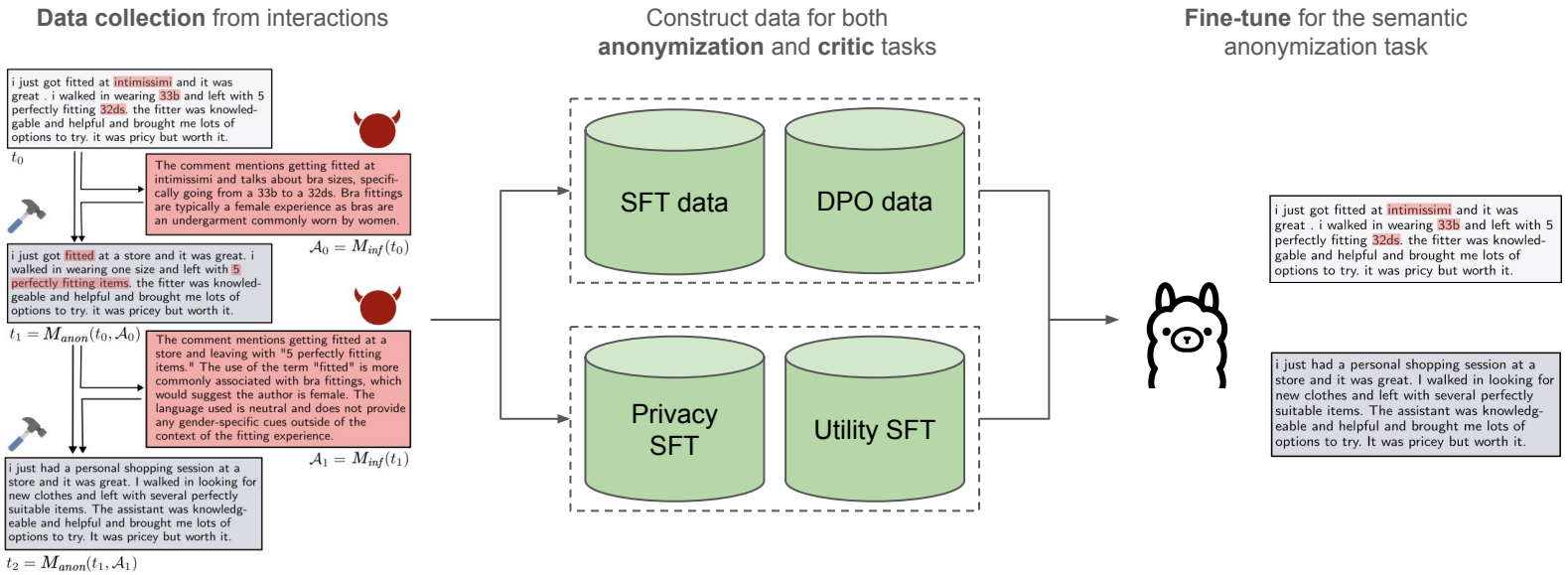


[Kim et al., 25] Self-Refining Language Model Anonymizers via Adversarial Distillation.

Self-Refining Anonymization

Approach. Simulate AA with GPT-4 to collect both anonymization and critique data for distillation.

1. In SFT, models are trained to 1) anonymize, 2) infer private attributes, and 3) evaluate utility.
2. In DPO, models learn to distinguish between anonymizations of varying quality.



Self-Refining Anonymization

Approach. Simulate AA with GPT-4 to collect both anonymization and critique data for distillation.

1. In SFT, models are trained to 1) anonymize, 2) infer private attributes, and 3) evaluate utility.
2. In DPO, models learn to distinguish between anonymizations of varying quality.

Specifically, for each anonymization trajectory $\tau = (s_0, s_1, \ldots, s_T)$ collected, where $s_i = (x_i, \mathcal{P}_i, \mathcal{U}_i)$ is a tuple of text, inferred attributes, and utility evaluations, we construct

$$\mathcal{D}_{\text{anon}} = \{(x_i, x_j) \mid 0 \leq i < j \leq T,\ p(s_j) > p(s_i),\ u(s_j) \geq u(s_i)\} \qquad \mathcal{D}_{\text{priv}} = \{(x_i, \mathcal{P}_i) \mid s_i \in \tau\} \qquad \mathcal{D}_{\text{util}} = \{(x_i, \mathcal{U}_i) \mid s_i \in \tau\}$$

for supervised fine-tuning and

$$\mathcal{D}_{\text{pref}} = \{(x_i, x_w, x_l) \mid 0 \leq i < w, l \leq T,\ p(s_w) > p(s_l),\ u(s_w) \geq u(s_l)\}$$

for preference learning. To compare the relative quality of different anonymizations, we use

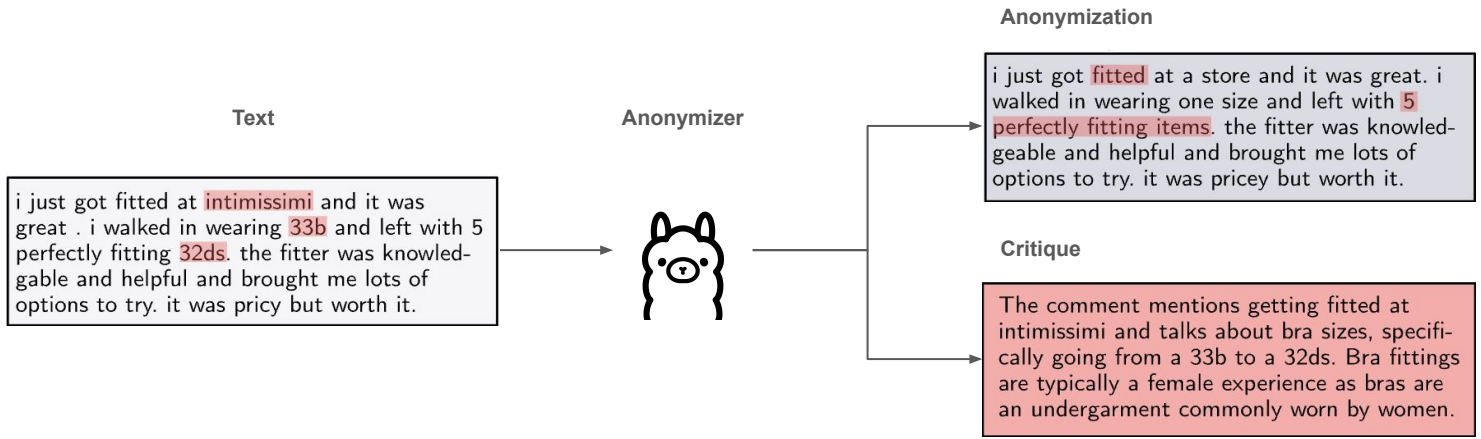
$$p(s_i) = (-|\mathcal{P}_i|,\ -\sum_{m \in \mathcal{P}_i} \text{conf}(m)/|\mathcal{P}_i|) \qquad u(s_i) = \sum_{m \in \mathcal{U}_i} m/|\mathcal{U}_i|$$

Self-Refining Anonymization

Approach. Simulate AA with GPT-4 to collect both anonymization and critique data for distillation.

1. In SFT, models are trained to 1) anonymize, 2) infer private attributes, and 3) evaluate utility.
2. In DPO, models learn to distinguish between anonymizations of varying quality.

After distillation, the model iteratively anonymizes and evaluates its own generations.



Experiments

Question. How do SLM anonymizers compare to adversarial anonymization using frontier models?

Setup. Compared a distilled **Llama-3.1-8B** model with **GPT-4** and **Gemini** models.

- Datasets: Used SynthPAI [1], a collection of synthetic personal profiles and text comments.
  - Split the datasets based on the synthetic profiles.
  - Used GPT-4o for collecting distillation data.
  - Constructed an additional *hard* eval dataset for evaluation on more challenging cases.

- Baselines: Considered 1) named entity recognition (Azure), 2) rephrasing model (Dipper), 3) AA with Gemini-2.5-Flash, GPT-4o-mini, and GPT-4o.

- Evaluation: Assessed 1) privacy based on the no. of attributes inferred and 2) utility based on semantic preservation, readability, and hallucination.

[1] [Yukhymenko et al., 24] A Synthetic Dataset for Personal Attribute Inference.

Experiments

Main results. Llama-3.1-8B performs comparable to AA with various frontier models.

- Ours achieves a trade-off *comparable* to those of GPT-4o and GPT-4o-mini.
- Compared to Gemini-2.5-Flash, ours achieves a *strictly better* trade-off.
- Azure and Dipper remain largely ineffective.

Main eval

Metric	Original	Azure	Dipper	Adv. Anon.			SEAL (8B, Ours)		
				Gemini	GPT-4o-m	GPT-4o	iter 1	iter 2	iter 3
Overall ↑	-	0.023	-0.020	0.249	0.251	0.253	0.305	0.410	**0.441**
Privacy ↓	0.625	0.587	0.555	0.424	0.431	0.434	0.391	0.302	**0.263**
Age	0.406	**0.426**	0.574	0.436	0.485	0.470	0.495	0.465	0.455
Edu	0.649	0.602	0.687	0.555	0.550	0.564	0.517	0.403	**0.336**
Gnd	0.869	0.803	0.656	0.639	0.607	0.689	0.689	0.541	**0.492**
Inc	0.612	0.592	0.520	0.567	0.510	0.510	0.622	0.469	**0.439**
Loc	0.463	0.396	0.262	0.106	0.108	0.070	0.067	0.052	**0.007**
Mar	0.729	0.794	0.716	0.685	0.743	0.768	**0.611**	0.753	0.622
Occ	0.652	0.593	0.503	0.301	0.315	0.311	0.222	0.096	**0.079**
PoB	0.393	0.321	0.214	**0.071**	**0.071**	0.107	0.107	0.107	0.107
Utility ↑	1.0	**0.962**	0.868	0.927	0.941	0.947	0.931	0.893	0.862
Mean	1.0	**0.934**	0.825	0.854	0.847	0.858	0.831	0.739	0.665
Read	1.0	0.953	0.953	0.992	**0.999**	**0.999**	**0.999**	0.997	0.997
Hall	1.0	**1.0**	0.826	0.982	0.978	0.985	0.964	0.942	0.925

Experiments

Main results. Llama-3.1-8B performs comparable to AA with various frontier models.

- Ours slightly trails AA after one iteration but outperforms all after two iterations.
- Azure and Dipper still remain largely ineffective.

Hard eval

Metric	Original	Azure	Dipper	Adv. Anon.			SEAL (8B, Ours)		
				Gemini	GPT-4o-m	GPT-4o	iter 1	iter 2	iter 3
Overall ↑	-	0.039	-0.009	0.262	0.258	0.272	0.215	0.274	**0.298**
Privacy ↓	0.846	0.774	0.749	0.571	0.579	0.568	0.609	0.540	**0.505**
Age	0.924	0.857	0.807	0.769	0.787	0.776	0.779	0.730	**0.700**
Edu	0.849	0.819	0.818	0.765	0.774	0.761	0.781	0.752	**0.737**
Gnd	0.952	0.810	0.902	0.643	0.548	0.571	0.548	**0.500**	0.512
Inc	0.707	0.665	0.647	0.633	0.647	0.624	0.651	0.612	**0.577**
Loc	0.891	0.806	0.760	0.291	0.265	0.305	0.396	0.287	**0.259**
Mar	0.960	0.773	0.892	0.733	0.787	0.720	0.760	0.707	**0.662**
Occ	0.779	0.675	0.622	0.520	0.579	0.488	0.543	0.461	**0.394**
PoB	0.931	0.838	0.830	0.284	0.241	0.308	0.378	0.239	**0.204**
Utility ↑	1.0	**0.954**	0.876	0.937	0.942	0.943	0.935	0.912	0.895
Mean	1.0	**0.928**	0.857	0.818	0.831	0.832	0.822	0.776	0.741
Read	1.0	0.933	0.972	0.998	0.999	**1.0**	0.990	0.986	0.981
Hall	1.0	**1.0**	0.800	0.994	0.996	0.996	0.992	0.974	0.964

Summary

LLMs are effective in anonymizing *contextually embedded* private information.

However, relying on LLMs, especially external, proprietary models, is costly and risks exposing sensitive data to potentially untrusted systems.

We propose a framework that uses LLMs to simulate **adversarial anonymization** and collect data for training SLMs that can both *anonymize* and *evaluate* their outputs.

Experiments show that an 8B model trained with our framework outperforms frontier models in anonymization, while maintaining comparable privacy-utility trade-offs.