Look Twice before You Leap: A Rational Framework for Localized Adversarial Text Anonymization

Donghang Duan[1], **Xu Zheng**[1*], **Yuefeng He**[1],
Chong Mu[1], **Leyi Cai**[2], **Lizong Zhang**[1]
[1]School of Computer Science and Engineering,
University of Electronic Science and Technology of China
[2]School of Information, Renmin University of China
xzheng@uestc.edu.cn

Abstract

Current LLM-based frameworks for text anonymization usually rely on remote API services from powerful LLMs, which creates an inherent privacy paradox: users must disclose the raw data to untrusted third parties for guaranteed privacy preservation. Moreover, directly migrating current solutions to local small-scale models (LSMs) offers a suboptimal solution with severe utility collapse. Our work argues that this failure stems not merely from the capability deficits of LSMs, but significantly from the inherent irrationality of the greedy adversarial strategies employed by current state-of-the-art (SOTA) methods. To address this drawback, we propose Rational Localized Adversarial Anonymization (RLAA), a fully localized and training-free framework featuring an Attacker-Arbitrator-Anonymizer architecture. We model the anonymization process as a trade-off between Marginal Privacy Gain (MPG) and Marginal Utility Cost (MUC), demonstrating that greedy strategies tend to drift into an irrational state. Instead, RLAA introduces an arbitrator that acts as a rationality gatekeeper, validating the attacker's inference to filter out ghost leaks. This mechanism promotes a rational early-stopping criterion, and structurally prevents utility collapse. Extensive experiments on different benchmarks demonstrate that RLAA achieves a superior privacy-utility trade-off compared to strong baselines.

1 Introduction

Large language models (LLMs) are increasingly deployed to process real-world text containing sensitive Personal Identifiable Information (PII), such as medical records, legal documents and online self-disclosures (Bommasani, 2021; Weidinger et al., 2021; Deußer et al., 2025). To comply with regulatory requirements like GDPR and CCPA (Albrecht, 2016; Bonta, 2022), effective text anonymization has become a prerequisite for the responsible use of such data. However, this situation presents a tricky trade-off between semantic utility and privacy, where over-anonymization destroys the semantic value while under-anonymization invites re-identification risks, which is particularly acute in sensitive domains like smart healthcare and legal technology (Im et al., 2024; Morris et al., 2025; Liu et al., 2025). The rapid evolution of LLMs has exacerbated this situation because of their dual roles as context-aware anonymizers and superior attackers (Wang et al., 2025). As a result, traditional NER-based anonymization methods are increasingly inadequate (Dernoncourt et al., 2017), motivating a shift toward semantic anonymization that conceals latent cues of identity beyond surface-level entities (Loiseau et al., 2025).

To address this situation, advanced anonymization frameworks based on LLMs have emerged. As the current state-of-the-art paradigm, Feedback-guided Adversarial Anonymization (FgAA) (Staab et al., 2024a) employs a dynamic adversarial game: an attacker model attempts to re-identify personal information from the text generated by an anonymizer model, and the anonymizer uses the inference feedback to instruct the anonymization process for refinement. Despite its strong empirical performance, FgAA heavily relies on the capabilities of powerful or closed-source LLMs like GPT-4 (Achiam et al., 2023), which are typically accessible only via third-party APIs. This dependency gives rise to a fundamental privacy paradox: to anonymize sensitive data, users must first disclose the raw content to external and uncontrollable service providers, which renders it unacceptable in high-sensitivity scenarios (Feretzakis and Verykios, 2024). Although subsequent works aim to mitigate this risk through localized deployment, they face distinct limitations. IncogniText (Frikha et al., 2024) compromises semantic utility by injecting ungrounded hallucinations. SEAL (Kim et al., 2025) necessitates complex training pipelines that

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 5556–5573

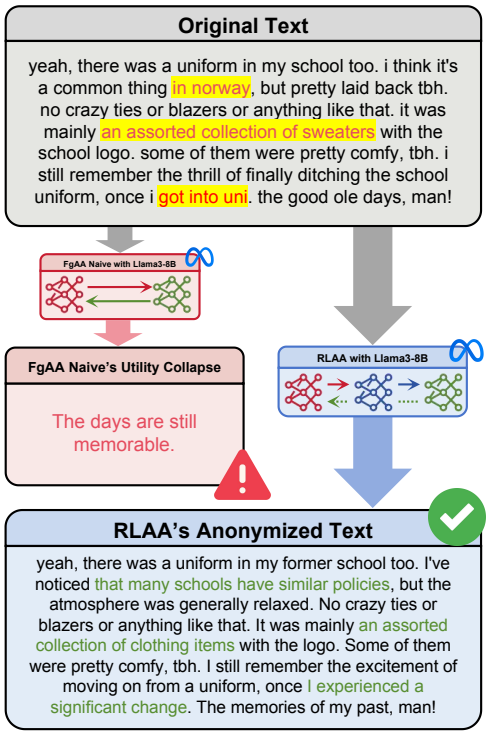

Figure 1: Utility Collapse of FgAA's Naive Migration.

rely on high-quality synthetic datasets which remain scarce (Yukhymenko et al., 2024). Besides, our empirical analysis in Section 4.3 reveals that it still suffers from severe utility collapse.

A natural and simple solution to the privacy paradox is to migrate adversarial anonymization frameworks to fully localized environments using local small-scale models (LSMs). However, our empirical investigation reveals that such a naive migration leads to severe utility collapse. As illustrated in Figure 1, the naive migration on Llama3-8B causes destructive over-editing. It indiscriminately strips away non-sensitive context and stylistic nuances, reducing the expressive original narrative to a generic and vacuous summary. Besides, it is observed that this collapse persists even when implementing instruction refinement (*e.g.*, outputting "unknown" for inferences with low confidence). This observation demonstrates that heuristic constraints alone cannot prevent the utility collapse in greedy adversarial strategies.

We argue that this failure cannot be solely attributed to the limited capabilities of LSMs, but instead arises from how greedy adversarial strategies behave under imperfect inferences. From an economic perspective, the anonymization can be viewed as a sequence of decisions balancing **Marginal Privacy Gain (MPG)** against **Marginal Utility Cost (MUC)**. By quantifying this ratio as the **Marginal Rate of Substitution (MRS** $= \frac{\text{MUC}}{\text{MPG}}$**)** (Mas-Colell et al., 1995), the primary driver of the utility collapse becomes clear through this economic lens: Current greedy strategies tend to drift into an economically inefficient state due to **LSMs' hallucinations** and **the law of diminishing returns**, where the model behaves irrationally by exhibiting exaggerated defense (Röttger et al., 2024) against **ghost leaks** that are either hallucinated or negligible (detailed in Appendix A). In this case, MUC remains positive while MPG approaches zero, pushing MRS to inefficient levels.

To overcome both the data dependency of distillation-based approaches and the utility collapse of naive migration, we propose **RLAA (Rational Localized Adversarial Anonymization)**[1], a training-free framework designed for fully localized deployment. The core idea of RLAA is to constrain the anonymization process to the economically rational condition defined in Section 3.2, thereby compensating for the capability deficits of LSMs without relying on explicit numerical optimization or parameter fine-tuning. RLAA introduces a novel **Attacker–Arbitrator–Anonymizer (A–A–A)** architecture, in which the arbitrator acts as a rationality gatekeeper between attack feedback and anonymization actions. Moving beyond the paradigm of blindly leaping into modifications, RLAA compels the arbitrator to validate the attacker's inferences, which structurally prevents utility collapse by rejecting destructive edits driven by ghost leaks, and maintains robust privacy protection by focusing on valid leaks. Crucially, this design leverages the **cognitive asymmetry** that verification is less complex and more reliable than generation for LSMs, which may hallucinate during open-ended inference but retain the ability to recognize errors in structured discrimination tasks (Cobbe et al., 2021; Wang et al., 2022; Madaan et al., 2023; Guan et al., 2024).

We conducted a comprehensive evaluation including baseline comparisons, ablation studies and generalization stress tests across several mainstream LLMs (Llama3-8B (AI@Meta, 2024), Qwen2.5-7B (Team, 2024) and DeepSeek-V3.2-Exp (DeepSeek-AI, 2025)) on the PersonalReddit and reddit-self-disclosure datasets. Additionally, we validated the framework's economic rationality through a quantitative analysis based on the MRS metric. The results empirically demonstrate that

[1]Our code and datasets are available at https://github.com/SowingG2333/RLAA.

RLAA structurally prevents utility collapse and achieves a superior privacy-utility trade-off compared to strong baselines. In summary, the main contributions of this study are as follows:

- We argue that migrating adversarial anonymization to local environments is crucial for eliminating the privacy paradox, while a naive migration results in utility collapse, identified as a symptom of the economic irrationality inherent in greedy strategies, rather than merely capability deficits.
- RLAA is proposed as a localized and training-free framework featuring an Attacker-Arbitrator-Anonymizer (A-A-A) architecture. By introducing an arbitrator, it structurally promotes rational decision-making to prevent utility collapse, providing a structured mechanism to preserve utility while reducing privacy risks without fine-tuning.
- Extensive experiments demonstrate that RLAA achieves a superior privacy-utility trade-off compared to multiple strong baselines and even Pareto dominance on the reddit-self-disclosure dataset, while also proving its mechanism generalization across different base models.

2 Related Work

2.1 Inference Attacks through LLMs

With increased reasoning capabilities, LLMs have evolved into potent privacy attackers capable of automatically inferring personal attributes from unstructured text (Wang et al., 2025). Staab et al. (2024b) demonstrated that models can accurately deduce fine-grained PII from casual online comments. Extending to high-stakes domains, Nyffenegger et al. (2024) revealed that LLMs could re-identify individuals in court decisions. Recent audits by Panda et al. (2025) further confirmed that this inference capability is consistent across various model architectures. These studies show that LLMs significantly lower the threshold for privacy breaches, enabling re-identification attacks to occur on an unprecedented scale (Staab et al., 2024b).

2.2 Text Anonymization

Recent advancements in text anonymization have marked a fundamental evolution from early statistical obfuscation techniques toward sophisticated semantic rewriting mechanisms. Generally, these studies can be divided into two categories:

Traditional Approaches. In early efforts, Differential Privacy (DP) methods (Abadi et al., 2016; Wu et al., 2023; Igamberdiev and Habernal, 2023) and representation learning (Coavoux et al., 2018; Friedrich et al., 2019) served as the standard. However, these methods often compromise textual readability or require training specific encoders, which limits their applicability to open-ended generation tasks. To address utility preservation, Shetty et al. (2018) introduced A4NT to obfuscate author attributes, while Mosallanezhad et al. (2019) utilized deep reinforcement learning to optimize the privacy-utility trade-off. These works laid the foundational mechanism for modern anonymizers.

LLM-driven Adversarial Frameworks. With the rise of generative AI, **Feedback-guided Adversarial Anonymization (FgAA)** (Staab et al., 2024a) established the prevailing paradigm for text anonymization. It systematizes the LLM's duality, leveraging an attacker model to iteratively critique and refine the anonymizer's output. Following this paradigm, **SEAL** (Kim et al., 2025) attempts to distill these capabilities into smaller models through SFT and DPO from multi-round teacher trajectories. However, this mechanism remains constrained by synthetic data scarcity (Yukhymenko et al., 2024) and potential generalization issues. **IncogniText** (Frikha et al., 2024) employs a misleading strategy, injecting false attributes to randomize attacker inferences rather than simply removing PII. Beyond general-purpose anonymization, **RUPTA** (Yang et al., 2025) addresses a complementary problem: optimizing anonymized text for specific downstream tasks (*e.g.*, occupation classification) using specialized task evaluators. While orthogonal to our focus on general semantic preservation, RUPTA shares the common reliance on API-based LLMs. Crucially, a common limitation across these advanced frameworks is their reliance on strong LLMs via APIs or extensive training pipelines. Moreover, methods following the greedy adversarial paradigm lack rational decision-making mechanisms, making them prone to utility collapse when deployed on smaller local models.

3 Methodology

3.1 Threat Model

RLAA is designed to defend against two distinct adversaries in the text anonymization task:

Semi-honest Service Provider ($\mathcal{A}_{serv}$). This adversary represents a third-party entity that hosts remote LLM services. It causes the privacy paradox: to utilize superior anonymizers, users must first transmit raw sensitive text x^{ori} to the provider.

This transmission exposes users to risks such as unauthorized data retention for model training or commercial profiling (Smith et al., 2023). Distinct from probable inference attacks, we characterize this threat as a deterministic data exposure, where $\mathcal{A}_{serv}$ gains full access to x^{ori} upon submission.

Re-identification Adversary ($\mathcal{A}_{re-id}$). This adversary represents any entity with access to the anonymized output x^{ano}. We assume $\mathcal{A}_{re-id}$ employs a powerful LLM $\mathcal{M}_{atk}$ (*e.g.*, DeepSeek-V3.2-Exp in our evaluation) to infer PII. The adversary's goal is to maximize the accuracy of the inferred attribute-value tuples $\{(a_j, v_j)\}$, where a_j denotes a specific private attribute category (*e.g.*, location) and v_j represents its corresponding inferred value (*e.g.*, "Paris, France"):

$$\{(a_j, v_j)\} \leftarrow \mathcal{M}_{atk}(x^{ano}) \tag{1}$$

3.2 Problem Formulation

The text anonymization task can be formalized as a constrained transformation problem. Let $\mathcal{X}$ be the space of all text sequences. Given an original text $x^{(0)} \in \mathcal{X}$, let $\mathcal{A}_{priv} = \{(a_k, v_k^*)\}_{k=1}^K$ be the set of "private attribute-true value" pairs contained in $x^{(0)}$, where a_k represents the attribute category (*e.g.*, age, location) and v_k^* denotes its true value. Let $\mathcal{M}_{atk}$ be a re-identification adversary. We define the privacy compromise metric $P_{asr}(x)$ as the adversary's success rate in inferring $\mathcal{A}_{priv}$ from a text x:

$$P_{asr}(x) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\left(\hat{v}_k \approx v_k^* \mid (a_k, \hat{v}_k) \in \mathcal{M}_{atk}(x)\right) \tag{2}$$

where $\mathbb{I}(\cdot)$ is the indicator function checking if the inferred value $\hat{v}_k$ semantically aligns with the ground truth v_k^*. This fuzzy matching operator ($\approx$) accounts for the inherent linguistic variability in LLM-generated inferences, ensuring the evaluation captures semantic equivalence rather than strict string identity.

Simultaneously, let $U(x^{(0)}, x) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a utility function measuring semantic preservation, where a higher value indicates better utility. The ideal objective is to find an optimum x^* such that privacy is protected while utility is maximized:

$$\max \quad U(x^{(0)}, x) \quad \text{s.t.} \quad P_{asr}(x) \leq \delta \tag{3}$$

where δ denotes the upper bound on acceptable privacy risk.

Migrating the framework to a local environment naturally eliminates the service provider threat $\mathcal{A}_{serv}$. However, achieving the objective in Eq. 3 is non-trivial. Existing adversarial frameworks typically approximate the optimum x^* via an iterative sequence $x^{(0)} \to x^{(1)} \cdots \to x^{(T)}$, where each step attempts to reduce privacy risk. Our empirical investigation reveals a critical challenge: directly applying greedy adversarial iterations with LSMs in this process often leads to utility collapse. We hypothesize that this failure arises because naive greedy strategies lack a mechanism to evaluate the cost-effectiveness of privacy gains. To formally address this, we reframe this iterative process over steps $t = 1, 2, \ldots T$ through an economic lens, viewing each anonymization operation as a transaction. The following marginal metrics are defined to quantify the rationality of each update.

Definition 1 (Marginal Privacy Gain, MPG). The reduction in adversarial inference accuracy achieved by the transformation at step t:

$$\Delta P_t = P_{asr}(x^{(t-1)}) - P_{asr}(x^{(t)}) \tag{4}$$

where $x^{(t-1)}$ represents the intermediate text generated at the previous iteration.

Definition 2 (Marginal Utility Cost, MUC). The semantic loss incurred at step t:

$$\Delta C_t = U(x^{(0)}, x^{(t-1)}) - U(x^{(0)}, x^{(t)}) \tag{5}$$

Definition 3 (Marginal Rate of Substitution, MRS). The instantaneous price of privacy preservation, representing the utility cost per unit of privacy gained:

$$\text{MRS}_t = \frac{\Delta C_t}{\Delta P_t} \tag{6}$$

From this perspective, a rational framework should ideally align with the principle that $\text{MRS}_t \leq \lambda$, where λ represents the maximum rational utility cost per unit of privacy. Naive greedy strategies inherently tend to violate this condition and drift into irrational states of deadweight loss where $\text{MRS}_t \to \infty$. Guided by this insight, RLAA is proposed to implicitly enforce this budget constraint λ through architectural rationality, thereby structurally preventing utility collapse. Detailed theoretical analysis is shown in Appendix A.

3.3 The RLAA Framework

To address the problem mentioned above, we propose the Attacker-Arbitrator-Anonymizer (A-A-A) architecture illustrated in Figure 2, which operates iteratively to refine the text $x^{(t)}$. For detailed algorithm settings and pseudo-code, please refer to Appendix C.1.

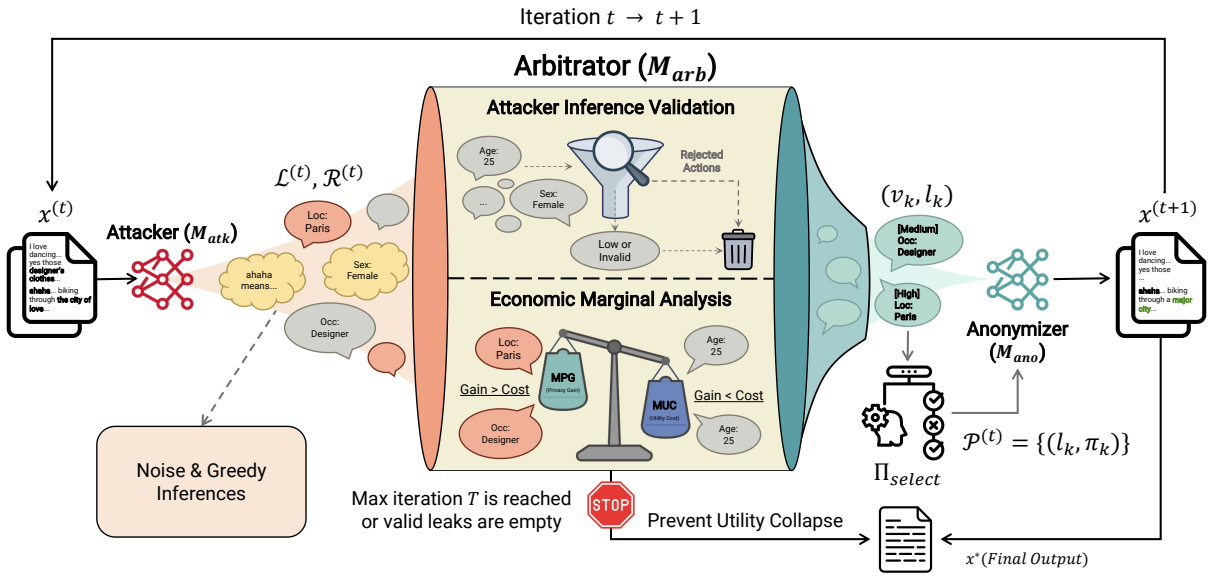

Figure 2: **The RLAA Framework.** Utilizing an Attacker-Arbitrator-Anonymizer architecture, the arbitrator acts as a rationality gatekeeper. It validates attacker inferences to filter out ghost leaks with negligible privacy benefits, structurally preventing utility collapse caused by irrational greedy strategies.

3.3.1 The Attacker (M_{atk})

The attacker acts as the sensory module. Given the current text $x^{(t)}$, it infers potential PII attributes and provides a reasoning chain:

$$(\mathcal{L}^{(t)}, \mathcal{R}^{(t)}) = M_{atk}(x^{(t)}) \tag{7}$$

where $\mathcal{L}^{(t)}$ is the set of identified leaks and $\mathcal{R}^{(t)}$ is the set of corresponding inferences.

3.3.2 The Arbitrator (M_{arb})

The arbitrator functions as the central control module that regulates anonymization decisions considering the rationality constraints defined in Section 3.2. It includes a generative LSM backbone for logic validation and a deterministic control layer that parses validity signals and enforces the filtering process. Instead of explicitly optimizing marginal trade-offs, which is unreliable given LSM's limited sensitivity to fine-grained scalar signals (Sun et al., 2025), the arbitrator validates attacker inferences by constraining the LSM to a structured discrimination task. Leveraging the greater reliability of verification over generation (Guan et al., 2024), this design enables self-correction and blocks ghost leaks from driving utility collapse.

For each leak $l_k \in \mathcal{L}^{(t)}$, the arbitrator assigns a validity level $v_k \in \mathcal{V}$, where the full validity space is $\mathcal{V} = \{\text{HIGH}, \text{MED}, \text{LOW}, \text{INVALID}\}$. From the economic perspective defined in Section 3.2, this validity level acts as a discrete estimator of MPG. We logically partition $\mathcal{V}$ into the valid set $\mathcal{V}_{valid}$ (representing $\Delta P_t > 0$) and the ghost set $\mathcal{V}_{ghost}$ (representing $\Delta P_t \approx 0$), and the detailed partition is provided in Appendix C.1. Accordingly, we define the selection policy Π_{select} as:

$$\Pi_{select}(v_k) = \begin{cases} \text{EXECUTE} & \text{if } v_k \in \mathcal{V}_{valid} \\ \text{IGNORE} & \text{if } v_k \in \mathcal{V}_{ghost} \end{cases} \tag{8}$$

The EXECUTE branch captures leakage with significant returns, ensuring a finite MRS. Conversely, the IGNORE branch filters out ghost leaks with $\Delta P_t \rightarrow 0$. By rejecting these transactions, the arbitrator prevents the system from drifting into the MRS singularity identified in Eq. 6, thereby structurally prevents utility collapse.

3.3.3 The Anonymizer (M_{ano})

The anonymizer executes the refined policy $\mathcal{P}^{(t)}$ with the actionable leaks validated by the arbitrator:

$$\mathcal{P}^{(t)} = \{(l_k, \pi_k) \mid l_k \in \mathcal{L}^{(t)}, \Pi_{select}(v_k) = \text{EXECUTE}\} \tag{9}$$

Based on this set, the text update is governed by:

$$x^{(t+1)} = \begin{cases} M_{ano}(x^{(t)}, \mathcal{P}^{(t)}) & \text{if } \mathcal{P}^{(t)} \neq \emptyset \\ x^{(t)} & \text{if } \mathcal{P}^{(t)} = \emptyset \end{cases} \tag{10}$$

When $\mathcal{P}^{(t)}$ is empty, the system triggers the early stop branch. This step guarantees the iteration converges to the fixed point $x^{(t+1)} = x^{(t)}$, thereby preventing the utility collapse caused by diminishing returns.

4 Experiments

4.1 Experimental Setup

Datasets. Two benchmark datasets are employed to evaluate our approach. **PersonalReddit** (Staab

et al., 2024b) is a widely adopted dataset in research including IncogniText and RUPTA, which consists of 525 human-verified synthetic Reddit conversations annotated with 8 fine-grained private attributes (*e.g.*, age, gender and location). **reddit-self-disclosure** (Dou et al., 2024) serves as a real-world evaluation bed, from which 885 samples containing self-disclosures of health conditions are extracted. Since this dataset is inherently labeled, we utilize the original labels as ground truth and perform manual verification to ensure consistency.

Models. Our experiments involve two distinct model roles: a local model for the anonymization process and an external evaluation model for benchmarking privacy and utility. We employ **Llama3-8B** and **Qwen2.5-7B** as local backbones, both of which facilitate consumer-grade deployment by requiring only approximately 4GB VRAM under 4-bit quantization. To evaluate privacy, we employ the 685B **DeepSeek-V3.2-Exp** as a robust re-identification adversary ($\mathcal{A}_{re-id}$), following the threat model defined in Section 3.1. Unlike proprietary APIs (*e.g.*, GPT-4), which impose strict cost and access constraints, DeepSeek enables large-scale and strong attacks at negligible cost due to its open-source nature, making it particularly suitable for modeling an economically rational adversary operating under realistic resource constraints. However, to mitigate the risk of self-referential evaluation from relying on a single backbone, we additionally re-evaluate the local methods using GPT-4o as an independent evaluator. Detailed results are reported in Appendix D.2.

Baseline Methods. We compared the performance of RLAA against **FgAA** (Staab et al., 2024a) (including its **Naive** and **SFT** migration variants), **SEAL** (Kim et al., 2025), **IncogniText** (Frikha et al., 2024) and **DP-BART-PR+** (Igamberdiev and Habernal, 2023). The maximum adversarial iteration limit T was configured according to dataset complexity and the need to evaluate varying convergence horizons: we set $T = 10$ for the multi-attribute PersonalReddit and $T = 3$ for the single-attribute reddit-self-disclosure. Detailed configurations are provided in Appendix C.2 and C.4.

4.2 Evaluation Metrics

To evaluate the performance of all methods, we adopt the standardized privacy and utility evaluation protocols widely established in recent LLM-based anonymization research (Staab et al., 2024a; Frikha et al., 2024; Kim et al., 2025). Let $D = \{(x_i^{ori}, A_i)\}_{i=1}^N$ be the test dataset, where x_i^{ori} is the original text and A_i is the set of ground-truth private attributes. Let x_i^{ano} be the anonymized text generated by a given method.

Privacy (PRIV). The performance on privacy preservation is evaluated using the anonymization under LLM inference setting (Staab et al., 2024a). We use a powerful adversary model $\mathcal{M}_{atk}$ (DeepSeek-V3.2-Exp) to infer the set of attributes $A_i' = \mathcal{M}_{atk}(x_i^{ano})$ from the anonymized text. The score is the average attack success rate over all attributes K across all N samples, which employs a programmatic matcher detailed in Appendix C.5.

$$\text{PRIV} = \frac{1}{N \cdot K} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{I}(A'_{i,k} \approx A_{i,k}) \tag{11}$$

Utility (UTIL). The utility evaluation relies on an LLM-as-a-Judge $\mathcal{M}_{judge}$ (DeepSeek-V3.2-Exp) (Zhang et al., 2024) to assess semantic preservation via K distinct components. We denote the scalar score range as $[S_{\min}, S_{\max}]$: **(1) Readability (s_{read})** is a scalar score $\in [S_{\min}, S_{\max}]$ assessing if the text is understandable on its own, defined as $s_{\text{read},i} = \mathcal{M}_{judge}(x_i^{ano})$. **(2) Meaning ($s_{\text{mean}}$)** is a scalar score $\in [S_{\min}, S_{\max}]$ assessing if the core message is preserved, where $s_{\text{mean},i} = \mathcal{M}_{judge}(x_i^{ori}, x_i^{ano})$. **(3) Hallucinations ($s_{\text{hall}}$)** is a binary score $s_{\text{hall},i} = \mathbb{I}(\text{no new ungrounded info})$ The final score is the average of these K normalized components ($K = 3, S_{\min} = 1, S_{\max} = 10$):

$$\text{UTIL} = \frac{1}{K \cdot N} \sum_{i=1}^{N} \left(\frac{s_{\text{read},i}}{S_{\max}} + \frac{s_{\text{mean},i}}{S_{\max}} + s_{\text{hall},i} \right) \tag{12}$$

Traditional Structural Metrics. In addition to LLM-based **UTIL Score**, we report standard n-gram metrics to evaluate structural similarity: **ROUGE** denotes the ROUGE-L F1 Score (Lin, 2004) based on the longest common subsequence, and **BLEU** (Papineni et al., 2002) indicates the bilingual evaluation understudy score for precision.

4.3 Performance Analysis

Our analysis begins by establishing RLAA's empirical superiority against strong baselines in Section 4.3.1 and verifying the arbitrator's crucial role through ablation studies in Section 4.3.2. The empirical findings are then bridged with our theoretical model via an economic analysis in Section 4.3.3. Finally, a human study in addition to automatic evaluations is presented in Section 4.3.4.

	Method	Base Model	UTIL ↑	PRIV ↓	ROUGE ↑	BLEU ↑	MEAN ↑	READ ↑	HALL ↑
					`PersonalReddit`				
	Original Text	-	1.0000	0.4442	1.0000	1.0000	10.000	10.000	1.0000
Local Methods	DP-BART+	BART-Base	0.3470	0.2650	0.3925	0.2597	2.6610	7.5790	0.0170
	IncogniText	Llama3-8B	0.6330	**0.1230**	0.3499	0.2304	5.5040	**10.0000**	0.3470
	FgAA-Naive	Llama3-8B	0.7297	0.1948	0.2180	0.0533	3.5420	8.9330	0.9420
	RLAA	Llama3-8B	**0.8788**	0.2130	**0.5958**	**0.4251**	**7.0780**	9.8090	**0.9480**
API Required	SEAL	Llama3-8B	0.4642	0.1787	0.1205	0.0753	9.2645	1.5207	0.3140
	FgAA-SFT	Llama3-8B	0.9670	0.2940	0.9149	0.9389	9.2310	9.9340	0.9830
	FgAA-API	DeepSeek-V3.2-Exp	0.8264	0.2056	0.4649	0.2082	5.4380	9.3550	1.0000
					`reddit-self-disclosure`				
	Original Text	-	1.0000	0.4943	1.0000	1.0000	10.000	10.000	1.0000
Local Methods	DP-BART+	BART-Base	0.3999	0.3245	0.2565	0.1093	2.3698	8.6830	0.0943
	IncogniText	Llama3-8B	0.7755	0.1283	**0.8584**	**0.8781**	**7.0792**	**9.9585**	0.6226
	FgAA-Naive	Llama3-8B	0.8187	0.1591	0.4919	0.2805	5.3523	9.6629	**0.9545**
	RLAA	Llama3-8B	**0.8572**	**0.1136**	0.6016	0.4703	6.8939	9.6932	0.9129
API Required	SEAL	Llama3-8B	0.6303	0.0226	0.1213	0.0736	9.8377	2.2415	0.6830
	FgAA-SFT	Llama3-8B	0.9218	0.1925	0.7813	0.7333	8.0528	9.9019	0.9698
	FgAA-API	DeepSeek-V3.2-Exp	0.9118	0.1660	0.7573	0.6626	7.5210	9.8720	0.9960

Table 1: **Main Baseline Comparison.** Performance of RLAA against local anonymization baselines (DP-BART+, IncogniText and FgAA-Naive) and baselines that require external API access during training, supervision, or some stage of the pipeline (SEAL, FgAA-SFT and FgAA-API).

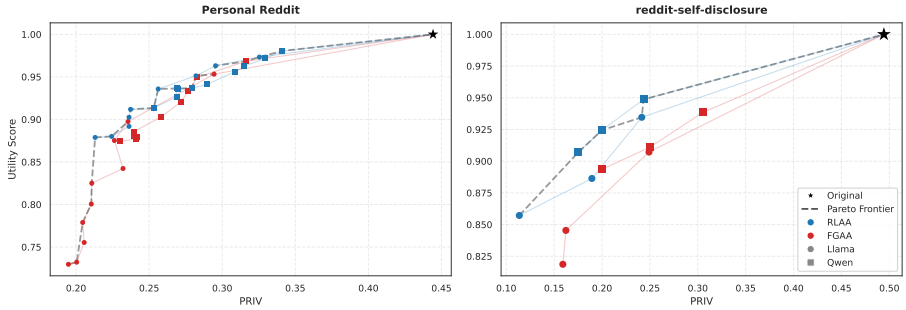

Figure 3: **Privacy-Utility Trade-off.** RLAA achieves superior trade-offs compared to FgAA across iterations on two datasets. The trade-off dynamics for structural metrics (ROUGE/BLEU) are detailed in Appendix D.1.

4.3.1 Baseline Comparison Results

Table 1 presents the comparative analysis on the PersonalReddit and reddit-self-disclosure datasets. RLAA achieves the optimal privacy-utility balance among local methods (UTIL=0.8788/0.8572). In contrast, FgAA-Naive experiences a pronounced utility collapse due to greedy over-editing. Figure 3 visualizes this performance gap across different horizons: FgAA exhibits progressive utility degradation consistent with diminishing returns on PersonalReddit, while RLAA maintains a superior Pareto frontier from the outset on reddit-self-disclosure. Besides, IncogniText sacrifices significant semantics (HALL=0.3470) by fabricating attributes that introduce ungrounded and potentially misleading contents. Distillation-based methods reveal a critical failure mode, *i.e.*, SEAL achieves strong privacy (PRIV=0.1787/0.0226) but catastrophic utility loss (UTIL=0.4642/0.6303). We argue this stems from distilling multi-round teacher trajectories without early-stopping signals: the student inherits privacy-seeking aggressiveness but never learns to stop. Conversely, our FgAA-SFT variant (detailed in Appendix D.4) corrects for over-caution (PRIV=0.2940/0.1925). It reveals distillation's fundamental limitation: it transfers editing behaviors but not the meta-judgment of rational stopping. In contrast, RLAA's training-free arbitrator structurally enforces this rationality, even rivaling FgAA-API (DeepSeek-685B) with 8B models.

4.3.2 Ablation Study

To isolate the arbitrator's contribution, we treat FgAA as the ablation baseline without arbitrator

Base Model	Method	UTIL ↑	PRIV ↓	ROUGE ↑	BLEU ↑	MEAN ↑	READ ↑	HALL ↑
				PersonalReddit				
Original Text	-	1.0000	0.4442	1.0000	1.0000	10.0000	10.0000	1.0000
DeepSeek-V3.2-Exp	w/o Arb.	0.8264	**0.2056**	0.4649	0.2082	5.4380	9.3550	**1.0000**
	RLAA	**0.9240**	0.2087	**0.7401**	**0.6365**	**7.8350**	**9.9670**	0.9920
Llama3-8B	w/o Arb.	0.7297	**0.1948**	0.2180	0.0533	3.5420	8.9330	0.9420
	RLAA	**0.8788**	0.2130	**0.5958**	**0.4251**	**7.0780**	**9.8090**	**0.9480**
Qwen2.5-7B	w/o Arb.	0.8741	**0.2302**	0.6156	0.3973	6.8590	9.6120	**0.9750**
	RLAA	**0.9135**	0.2531	**0.7549**	**0.6413**	**7.8020**	**9.8620**	0.9740
				reddit-self-disclosure				
Original Text	-	1.0000	0.4943	1.0000	1.0000	10.0000	10.0000	1.0000
DeepSeek-V3.2-Exp	w/o Arb.	0.9118	0.1660	0.7573	0.6626	7.5210	9.8720	**0.9960**
	RLAA	**0.9132**	**0.1434**	**0.8070**	**0.7364**	**7.5620**	**9.9090**	0.9920
Llama3-8B	w/o Arb.	0.8187	0.1591	0.4919	0.2805	5.3523	9.6629	**0.9545**
	RLAA	**0.8572**	**0.1136**	**0.6016**	**0.4703**	**6.8939**	**9.6932**	0.9129
Qwen2.5-7B	w/o Arb.	0.8936	0.2000	0.6458	0.4800	7.2115	9.8269	**0.9769**
	RLAA	**0.9071**	**0.1753**	**0.7273**	**0.6208**	**7.7200**	**9.9320**	0.9560

Table 2: **Ablation and Generalization**. Comparison between the full RLAA framework and the baseline without arbitrator. The arbitrator consistently improves utility across different model scales and datasets.

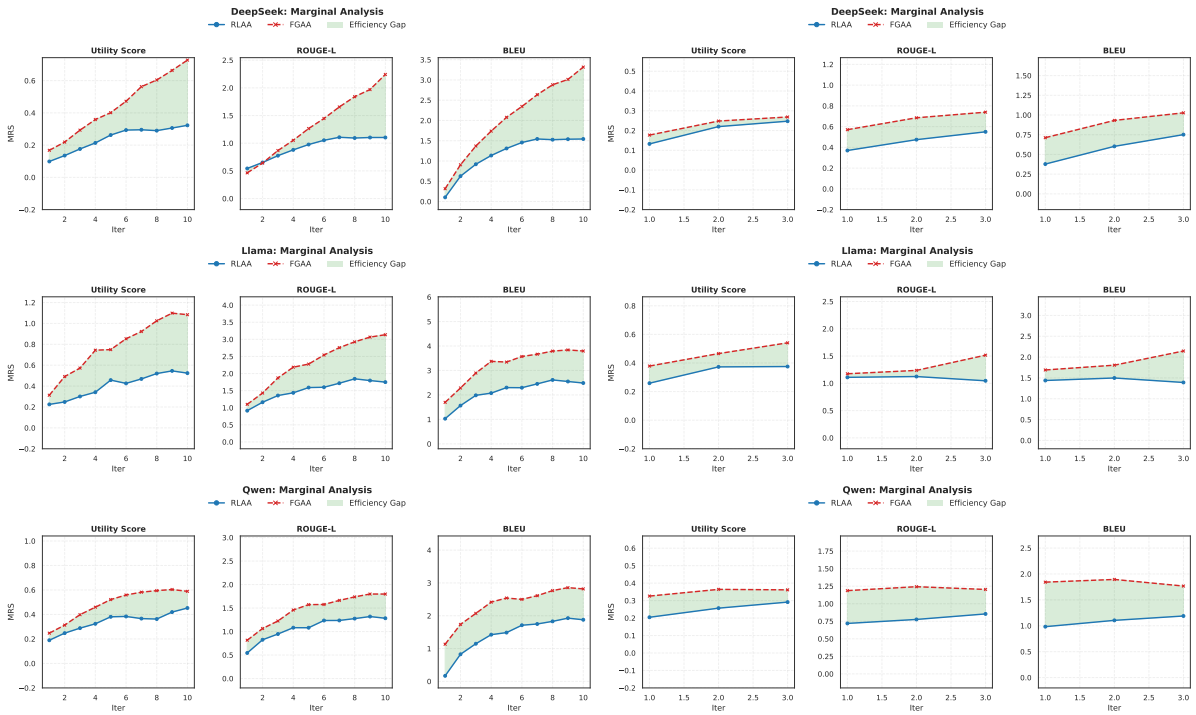

Figure 4: **Cumulative MRS Analysis.** The figure displays the cumulative MRS during the anonymization process on PersonalReddit (Left) and reddit-self-disclosure (Right). FgAA (Red) shows a sustained increase while RLAA (Blue) maintains a stable low MRS. Detailed quantitative economic analysis is provided in Appendix D.3.

w/o Arb. Table 2 shows consistent utility gains across all models. Crucially, RLAA achieves Pareto dominance on reddit-self-disclosure, improving both privacy (0.1591→0.1136) and utility (0.8187→0.8572), which demonstrates that rationality constraints actively optimize rather than merely constrain. Even the powerful DeepSeek-685B benefits from the arbitrator, which confirms again the irrationality also stems from the greedy strategy itself. This observation validates RLAA as a general component to compensate for the rationality of model behaviors, regardless of scales.

4.3.3 Economic Efficiency Analysis

To validate our theoretical premise regarding rationality, we analyze the marginal dynamics of the anonymization process. Figure 4 visualizes the MRS across iterations for Utility Score. FgAA

(Red Curve) exhibits a continuous increase in MRS across both datasets, which confirms that the greedy strategy tends to drift into an economically inefficient state. In contrast, RLAA (Blue Curve) maintains a low and stable MRS trajectory, reducing the terminal MRS based on UTIL Score from **3.80** to **1.74** on Llama3-8B. The comparison serves as an empirical proof of our hypothesis: The constraint on rationality prevents utility collapse. Besides, further quantitative analysis in Appendix D.3 reveals a counter-intuitive capability-rationality paradox: SOTA level models exhibit a steeper irrationality drift and achieve higher rationality gains than smaller models. This observation suggests that scaling capabilities alone cannot resolve rationality alignment failures.

4.3.4 Human Evaluation

To further examine the semantic utility of RLAA's generated text and the arbitrator's reliability, we conduct two complementary human evaluations on PersonalReddit dataset.

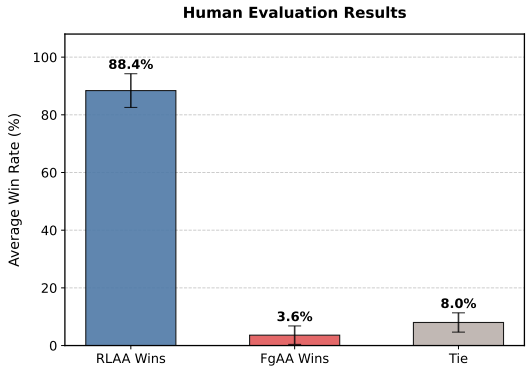

Figure 5: **Human pairwise evaluation results.** RLAA achieves a dominant win rate against FgAA-Naive while reflecting high inter-annotator consistency.

First, to directly assess semantic utility, we conduct a blind three-way pairwise comparison between the outputs of RLAA, FgAA-Naive and the original text. The evaluation involves five independent non-author annotators, each judging 50 randomized samples, resulting in 250 total judgments. As shown in Figure 5, RLAA achieved a dominant average win rate of **88.4%**, while the FgAA's win rate was only **3.6%**. RLAA's decisive preference rating substantiates that the arbitrator effectively preserves semantic integrity. We also observed a strong consensus across all five reviewers with a remarkably low standard deviation in win rates ($\sigma = 5.9\%$), which provides strong statistical evidence for inter-annotator consistency and ensures the objective reliability of these findings.

Metric	Llama3-8B	Qwen2.5-7B
Recall	95.45%	85.71%
Precision	58.33%	60.00%
Accuracy	84.00%	85.00%
FNR	4.55%	14.29%
FPR	19.23%	15.19%
F1	0.7241	0.7059

Table 3: **Human-validated arbitrator reliability.** FNR and FPR denote false negative and false positive rates.

Tier	Llama3-8B	Qwen2.5-7B
`HIGH`	90.00%	90.00%
`MED`	36.67%	53.33%
`LOW`	20.00%	10.00%
`INVALID`	3.33%	3.33%

Table 4: **Tier-based calibration.** Entries are the proportion of human-confirmed genuine leaks in each tier.

Second, to evaluate whether the arbitrator can reliably distinguish genuine leaks from ghost leaks, we conduct a human validation study on PersonalReddit with three annotators with NLP research backgrounds. The annotated set contains 200 randomly sampled instances from anonymization trajectories for overall reliability evaluation and 240 additional samples stratified across validity tiers for tier-based calibration; majority vote is used as the final label. Table 3 shows that the arbitrator achieves high recall and relatively low false-negative rates across both Llama3-8B and Qwen2.5-7B, indicating that genuine leaks are rarely filtered out by early stopping. To further assess calibration, we compute the proportion of human-confirmed genuine leaks within each validity tier. As shown in Table 4, this proportion decreases monotonically from `HIGH` to `INVALID` for both local backbones, supporting the interpretation that validity tiers provide an empirically grounded proxy for actionable privacy severity.

5 Conclusion

This paper reframes utility collapse in localized adversarial anonymization as a failure of rational decision-making and proposes RLAA, a training-free framework built on an Attacker-Arbitrator-Anonymizer architecture. By validating attacker inferences and introducing verification-guided early stopping, RLAA prevents semantic utility collapse caused by hallucinations and diminishing returns. Extensive experiments show that RLAA consistently achieves a stronger privacy-utility trade-off than competitive baselines. Grounding safety in economic rationality, RLAA resolves the privacy paradox, aligning safety with model capability.

Limitations

While RLAA achieves superior performance in extensive experiments, there are several limitations to acknowledge and to be mitigated in future work:
Computational Overhead. The current RLAA operates as an inference-time alignment mechanism, where the arbitrator enforces a slow-thinking verification pass. It is necessary for rationality but introduces additional computational overhead: A detailed quantitative analysis is provided in Appendix B, which indicates that this overhead remains acceptable for most quality-centric offline scenarios. Future work could leverage RLAA to generate high-quality and rationality-aligned trajectories. By fine-tuning models on these trajectories, we can internalize the arbitrator's external constraints into the model's parameters to achieve training-time alignment, thereby eliminating the computational overhead of an external gatekeeper.
Evolving Adversarial Capabilities. Our evaluation relies on DeepSeek-V3.2-Exp to simulate a powerful adversary. While this represents a current SOTA threat model, the landscape of LLM attacks is rapidly evolving. RLAA provides empirical defense, but we cannot theoretically guarantee immunity against future models with significantly stronger reasoning capabilities or attackers possessing extensive auxiliary background knowledge.
Lack of Provable Guarantees. Unlike Differential Privacy (DP) mechanisms, RLAA does not provide mathematically provable privacy budgets (ϵ). This is a known trade-off in LLM-based text anonymization: DP methods often result in significant degradation of semantic readability and utility. Our approach prioritizes semantic preservation and empirical safety, positioning it as a pragmatic solution rather than a mathematically guaranteed one.

Ethics Statement

We acknowledge the dual-use risks of LLMs and employ our adversarial framework strictly for defensive evaluation. Our approach enhances privacy by enabling fully localized deployment, thereby eliminating data exposure to third-party APIs. We clarify that RLAA provides empirical defense against SOTA adversaries, rather than provable guarantees like Differential Privacy. Regarding data usage, we rely exclusively on established public datasets (PersonalReddit and reddit-self-disclosure) and strictly adhere to ethical standards by avoiding any new collection of private data.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No.62372085, 62271125), the Sichuan Science and Technology Program (Grant No.2024NSFTD0033), and the China Postdoctoral Science Foundation (Grant No.2025M78146).

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Jan Philipp Albrecht. 2016. How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287.

Rishi Bommasani. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Rob Bonta. 2022. California consumer privacy act (ccpa). *Retrieved from State of California Department of Justice: https://oag. ca. gov/privacy/ccpa*.

Maximin Coavoux, Shashi Narayan, and Shay B Cohen. 2018. Privacy-preserving neural representations of text. *arXiv preprint arXiv:1808.09408*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Tobias Deußer, Lorenz Sparrenberg, Armin Berger, Max Hahnbück, Christian Bauckhage, and Rafet Sifa. 2025. A survey on current trends and recent advances in text anonymization. *arXiv preprint arXiv:2508.21587*.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754.

Georgios Feretzakis and Vassilios S Verykios. 2024. Trustworthy ai: Securing sensitive data in large language models. *AI*, 5(4):2773–2800.

Max Friedrich, Arne Köhn, Gregor Wiedemann, and Chris Biemann. 2019. Adversarial learning of privacy-preserving text representations for de-identification of medical records. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5829–5839.

Ahmed Frikha, Nassim Walha, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2024. Incognitext: Privacy-enhancing conditional text anonymization via llm-based private attribute randomization. *arXiv preprint arXiv:2407.02956*.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pages 1090–1111.

Timour Igamberdiev and Ivan Habernal. 2023. Dp-bart for privatized text rewriting under local differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13914–13934.

Eunyoung Im, Hyeoneui Kim, Hyungbok Lee, Xiaoqian Jiang, and Ju Han Kim. 2024. Exploring the tradeoff between data privacy and utility with a clinical data analysis use case. *BMC Medical Informatics and Decision Making*, 24(1):147.

Kyuyoung Kim, Hyunjun Jeon, and Jinwoo Shin. 2025. Self-refining language model anonymizers via adversarial distillation. *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Junkai Liu, Yujie Tong, Hui Huang, Bowen Zheng, Yiran Hu, Peicheng Wu, Chuan Xiao, Makoto Onizuka, Muyun Yang, and Shuyuan Zheng. 2025. Legal fact prediction: the missing piece in legal judgment prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6345–6360.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. 2025. Tau-eval: A unified evaluation framework for useful and private text anonymization. *arXiv preprint arXiv:2506.05979*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*, volume 1. Oxford university press New York.

John X Morris, Thomas R Campion, Sri Laasya Nutheti, Yifan Peng, Akhil Raj, Ramin Zabih, and Curtis L Cole. 2025. Diri: Adversarial patient reidentification with large language models for evaluating clinical text anonymization. *AMIA Summits on Translational Science Proceedings*, 2025:355.

Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2360–2369.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2024. Anonymity at risk? assessing re-identification capabilities of large language models in court decisions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2433–2462.

Srikant Panda, Hitesh Laxmichand Patel, Shahad Al-Khalifa, Amit Agarwal, Hend Al-Khalifa, and Sharefah Al-Ghamdi. 2025. Daiq: Auditing demographic attribute inference from question in llms. *arXiv preprint arXiv:2508.15830*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4NT: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1633–1650, Baltimore, MD. USENIX Association.

Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. 2023. Identifying and mitigating privacy risks stemming from language models: A survey. *arXiv preprint arXiv:2310.01424*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024a. Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846*.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024b. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Zhishen Sun, Guang Dai, Ivor Tsang, and Haishan Ye. 2025. Numerical sensitivity and robustness: Exploring the flaws of mathematical reasoning in large language models. *arXiv preprint arXiv:2511.08022*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Shang Wang, Tianqing Zhu, Bo Liu, Ming Ding, Dayong Ye, Wanlei Zhou, and Philip Yu. 2025. Unique security and privacy threats of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(4):1–36.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Tong Wu, Ashwinee Panda, Jiachen T Wang, and Prateek Mittal. 2023. Privacy-preserving in-context learning for large language models. *arXiv preprint arXiv:2305.01639*.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. 2024. A synthetic dataset for personal attribute inference. *Advances in Neural Information Processing Systems*, 37:120735–120779.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024. Small language models need strong verifiers to self-correct reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15637–15653.

Appendix

A Theoretical Analysis

In this section, we provide a formal economic definition and analytical derivation showing that adversarial strategies like FgAA are inherently irrational, while RLAA achieves structural rationality.

Definition 4 (**Economic Rationality Condition**)**.** An anonymization framework is defined as economically rational only if every executed step satisfies the budget constraint:

$$\text{MRS}_t \leq \lambda \tag{13}$$

where $\lambda \in \mathbb{R}^+$ represents the maximum acceptable utility cost per unit of privacy gain. λ is an intrinsic property derived from the specific utility metric $U(\cdot)$ and deployment constraints, representing the break-even point where the marginal cost of anonymization outweighs its privacy benefit.

Assumption 1 (**Discrete Estimation**)**.** We model the arbitrator's discrete validity judgment as a quantized estimator of the expected privacy gain $\mathbb{E}[\Delta P]$. Using the sets defined in Section 3.3 ($\mathcal{V}_{valid}$ and $\mathcal{V}_{ghost}$), we map discrete labels to expected gains:

$$\mathbb{E}[\Delta P_t(l)] \approx \begin{cases} \gamma_v > 0 & \text{if } v \in \mathcal{V}_{valid} \\ \gamma_g \approx 0 & \text{if } v \in \mathcal{V}_{ghost} \end{cases} \tag{14}$$

where γ_v represents a significant privacy improvement, and γ_g represents the negligible gain characteristic of a ghost leak.

Proposition 1 (**Implicit Budget Enforcement**)**.** *The arbitrator's discrete mechanism creates a binary decision boundary that is functionally equivalent to an economic mechanism operating with an implicit budget λ.*

Derivation. Assuming that the utility cost for any atomic edit is lower-bounded by $\Delta C_t \geq \epsilon$, we define the MRS lower bound for actionable leaks and the MRS upper bound for ghost leaks:

$$\text{MRS}_{valid} \approx \frac{\epsilon}{\gamma_v}, \quad \text{MRS}_{ghost} \approx \frac{\epsilon}{\gamma_g} \to \infty \tag{15}$$

Under the assumption of correct estimation, there exists a significant separation gap: $\text{MRS}_{valid} \ll \text{MRS}_{ghost}$. And the arbitrator's policy Π_{select} effectively implements a budget constraint λ located within this gap:

$$\begin{gathered} \exists \lambda \in \left[\frac{\epsilon}{\gamma_v}, \frac{\epsilon}{\gamma_g}\right) \\ \text{s.t. } \forall v, \Pi_{select}(v) \neq \textsc{Ignore} \iff \text{MRS}(v) \leq \lambda \end{gathered} \tag{16}$$

This mechanism ensures the system structurally prioritizes high-return transactions while rejecting deadweight losses caused by ghost leaks. □

Corollary 1 (Hallucination Defense - Instantaneous Rationality). *In a greedy system, hallucinated ghost leaks l_{hall} incur definitive utility costs (ϵ) for negligible privacy gains (γ_g). We can formally derive the irrationality of executing such edits as a singularity in the MRS:*

$$\lim_{\Delta P \to \gamma_g} MRS(l_{hall}) \approx \frac{\epsilon}{\gamma_g} \to \infty \tag{17}$$

*By applying **Proposition 1**, RLAA identifies such instances as belonging to the ghost set ($v \in \mathcal{V}_{ghost}$). Since the implicit MRS exceeds any rational budget λ, the policy triggers a rejection $\Pi_{select} =$ IGNORE. Thus, the transaction cost is forced to zero ($\Delta C_t = 0$), thereby avoiding the immediate deadweight loss.*

Corollary 2 (Diminishing Returns - Asymptotic Rationality). *As the iteration t increases, the process enters a long-tail phase where remaining candidates are exclusively ghost leaks with $\lim_{t\to\infty} \Delta P_t = \gamma_g$. A greedy strategy fails to stop because it lacks a mechanism to evaluate the diverging cost-benefit ratio:*

$$\lim_{t\to\infty} MRS_{greedy} = \frac{\Delta C_t}{\lim_{t\to\infty} \Delta P_t} \to \infty \tag{18}$$

*In contrast, RLAA applies **Proposition 1** to the entire set of remaining candidates $\mathcal{L}^{(t)}$. Since $\forall l \in \mathcal{L}^{(t)}, l \in \mathcal{V}_{ghost} \implies MRS(l) > \lambda$, the policy set becomes empty ($\mathcal{P}^{(t)} = \emptyset$). This explicitly triggers the algorithmic early stop, causing the system to converge to a stable rational equilibrium state $x^{(t)}$:*

$$x^{(t+1)} = x^{(t)} \quad (\textit{Stop Condition}) \tag{19}$$

This mechanism structurally counteracts the drift into utility collapse observed in naive baselines.

B Computational Overhead Analysis

To quantify the additional computation introduced by the arbitrator, we report both the inference latency and average token consumption under the same experimental configuration as Section 4.1 using Llama3-8B on an NVIDIA RTX 5090 GPU. As shown in Table 5, RLAA introduces moderate overhead on reddit-self-disclosure and larger overhead on PersonalReddit, which is expected given the latter's more complex multi-attribute reasoning setting. In addition, on PersonalReddit under a 10-iteration setting, RLAA increases the average token consumption from 18,229 to 31,178 tokens per sample ($\approx 1.71\times$) as shown in Table 6. This extra cost arises from the verification passes introduced by the arbitrator rather than from any increase in parameter memory, since the attacker, arbitrator and anonymizer share the same frozen local backbone. Given that anonymization is typically used as an offline or pre-deployment preprocessing step, we view this additional computation as an acceptable trade-off for improved stability and reduced over-editing.

Method	Inference Latency	Overhead
`reddit-self-disclosure`		
FgAA-Naive	14.7s / sample	1.00×
RLAA	21.6s / sample	1.47×
`PersonalReddit`		
FgAA-Naive	95.1s / sample	1.00×
RLAA	196.8s / sample	2.07×

Table 5: **Inference latency comparison.** RLAA incurs additional inference latency due to the verification passes introduced by the arbitrator.

Method	Avg. Tokens / Sample	Overhead
FgAA-Naive	18,229	1.00×
RLAA	31,178	1.71×

Table 6: **Total compute cost on PersonalReddit.** Average token consumption per sample under a 10-iteration setting. RLAA requires additional compute due to verification passes.

C Implementation Details

In this part, we provide implementation configurations, including the detailed algorithmic procedure, computational environment, training recipes and generation hyperparameters.

C.1 Algorithmic Procedure.

The detailed pseudo-code of RLAA is shown in **Algorithm 1** and we employ dataset-specific partitions for $\mathcal{V}_{valid}$ according to task complexity:

- **PersonalReddit**: $\mathcal{V}_{valid} = \{\text{HIGH}, \text{MED}\}$ to capture the multi-attribute and implicit identity cues inherent in this dataset.
- **reddit-self-disclosure**: $\mathcal{V}_{valid} = \{\text{HIGH}\}$ because health-issue disclosures are primarily explicit in this dataset.

This design highlights RLAA's flexibility: $\mathcal{V}_{valid}$ serves as a tunable hyperparameter, allowing the arbitration policy to align with specific economic constraint λ and domain-specific sensitivities: A stricter $\mathcal{V}_{valid}$ (e.g., $\{\text{HIGH}\}$) is preferred for explicit leaks to prioritize utility preservation, while

Algorithm 1 Rational Anonymization

Require: Original Text $x^{(0)}$, Max Iterations T
Ensure: Anonymized Text x^*
$t \leftarrow 0$
while $t < T$ **do**
  // Phase 1: Adversarial Inference
  $\mathcal{L}^{(t)}, \mathcal{R}^{(t)} \leftarrow \mathcal{M}_{atk}(x^{(t)})$
  // Phase 2: Rational Arbitration
  $\mathcal{P}^{(t)} \leftarrow \emptyset$
  for each pair (l_k, r_k) in $(\mathcal{L}^{(t)}, \mathcal{R}^{(t)})$ **do**
    $v_k \leftarrow \mathcal{M}_{arb}(l_k, r_k, x^{(t)})$
    $\pi_k \leftarrow \Pi_{select}(v_k)$
    if $\pi_k \neq$ IGNORE **then**
      $\mathcal{P}^{(t)} \leftarrow \mathcal{P}^{(t)} \cup \{(l_k, \pi_k)\}$
    end if
  end for
  // Phase 3: Execution & Early Stop
  if $\mathcal{P}^{(t)} = \emptyset$ **then break**
  end if
  $x^{(t+1)} \leftarrow \mathcal{M}_{ano}(x^{(t)}, \mathcal{P}^{(t)})$
  $t \leftarrow t + 1$
end while
return $x^{(t)}$

a looser setting (e.g., including {MED}) is recommended for handling nuanced stylistic identifiers. In practice, users can adapt the framework to various privacy-utility trade-offs by adjusting this discrete threshold without additional training.

To further examine the effect of this threshold choice, we conduct a sensitivity analysis on PersonalReddit with Llama3-8B by varying which validity tiers are treated as actionable leaks. The results are shown in Table 7. Using only `HIGH` preserves the most semantic utility but leaves substantially more residual leakage. In contrast, treating `HIGH`, `MED` and `LOW` all as actionable leaks reduces semantic utility while providing limited additional privacy benefit. The default setting used in the main experiments achieves the best balance, which supports our threshold choice for this dataset.

Policy	UTIL ↑	PRIV ↓
`HIGH` only	**0.9000**	0.2958
`HIGH+MED` (Default)	0.8788	**0.2130**
`HIGH+MED+LOW`	0.8369	0.2242

Table 7: **Threshold sensitivity of the arbitrator gate.** Results on PersonalReddit with Llama3-8B under different validity-threshold policies.

C.2 Base Models & Environment.

All local LLM-based frameworks (RLAA, FgAA, IncogniText) employed Llama-3-8B and Qwen2.5-7B as base models. To align with consumer-grade deployment scenarios, all models were loaded in half-precision (`float16`) on a single **NVIDIA RTX 5090 GPU**.

C.3 Training Configurations for Baselines.

Table 8 details the specific hyperparameter settings for all training-based baselines.

- **FgAA-SFT:** We performed standard supervised fine-tuning on Llama-3-8B for 10 epochs to ensure convergence. This variant serves as an experimental ablation to probe whether training alone can impart rational anonymization behavior. Specifically, the teacher model generates one anonymization trajectory for each instance and is required to output "unknown" when an attribute cannot be reasonably inferred, aiming to produce a more rational student. All LLM fine-tuning utilized QLoRA (4-bit) for memory efficiency.
- **SEAL:** For this distillation-based baseline, we strictly adhered to the official implementation, including its two-stage pipeline: an SFT phase followed by a conservative DPO phase, executed as specified by the released recipe.
- **DP-BART-PR+:** According to the official code, this baseline was trained with a gradient clipping norm $C = 5.0$ and a privacy budget $\epsilon = 2500$.

Param	FgAA-SFT	SEAL-SFT	SEAL-DPO	DP-BART
LR	1e−5	2e−4	5e−6	1e−5
Batch	4×1	4×2	4×1	16
Epoch	10	1	1	50
Max Len	1024	4096	2048	300
LoRA r	16	16	–	–
LoRA α	32	16	16	–
DPO β	–	–	0.01	–
DP ϵ	–	–	–	2500
DP δ	–	–	–	10^{-6}

Table 8: **Training hyperparameters for Baselines.** All models utilize 4-bit QLoRA to ensure efficiency.

Framework	Module	Temp	Top-p	Max Tokens
RLAA	Attacker	0.1	0.9	1024
	Arbitrator	0.0	-	1024
	Anonymizer	0.5	0.9	512
FgAA	Attacker	0.1	0.9	1024
	Anonymizer	0.5	0.9	512
IncogniText	Persona Gen.	0.7	1.0	512
	Local Anon.	0.5	0.9	512

Table 9: **Specific Generation Hyperparameters.** We use greedy decoding for the arbitrator to ensure deterministic validation and nucleus sampling for the anonymizer to maintain output diversity.

C.4 Generation Hyperparameters.

Table 9 details the generation hyperparameters for each module.

- **RLAA Settings:** The arbitrator uses a temperature of 0 to ensure deterministic validity judgments, while the attacker and anonymizer use a

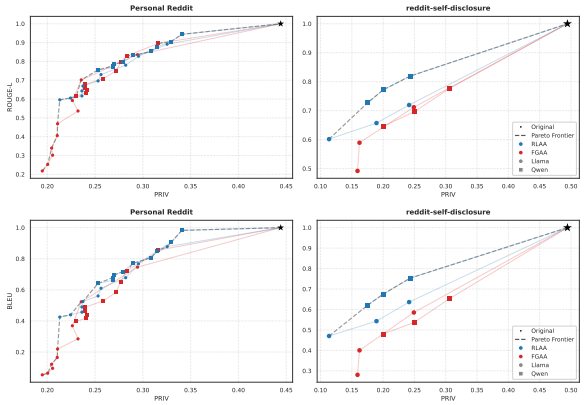

Figure 6: **Privacy-utility trade-offs via structural metrics (ROUGE and BLEU).** Results are shown for PersonalReddit (Left) and reddit-self-disclosure (Right), demonstrating RLAA's resistance to structural collapse.

slight temperature (0.1 and 0.5) to balance creativity and instruction following.

- **IncogniText Pipeline:** We adopt a two-stage pipeline where the target persona is pre-generated offline via DeepSeek-V3.2-Exp to encourage stylistic diversity. Crucially, API usage is strictly limited to this one-time data preparation phase, ensuring the subsequent anonymization process remains fully localized.

C.5 Privacy Evaluation Protocol.

To ensure evaluation objectivity, we calculate the PRIV score using a programmatic matcher instead of a stochastic LLM judge. By aligning the attacker's output schema with the evaluator through specific prompting, we enable rigorous programmatic verification:

- **Numerical Tolerance (Age)**: An attack is successful if $|\text{Age}_{true} - \text{Age}_{guess}| \leq 3$. This window accounts for the natural ambiguity in age inference from social media text.
- **Exact Match (Sex, Income, Status)**: We apply strict case-insensitive exact string matching.
- **Sub-string Match (Location, Job, Education, Health Issue)**: We employ bi-directional substring matching.

This protocol serves as a strict proxy for privacy risk, prioritizing high sensitivity to potential exposures while ensuring complete reproducibility.

D Detailed Experimental Results

This section expands on the main analysis by providing structural evaluations, evaluation results with GPT-4o as judge, training stability checks and a quantitative economic analysis that reveals a counter-intuitive alignment paradox.

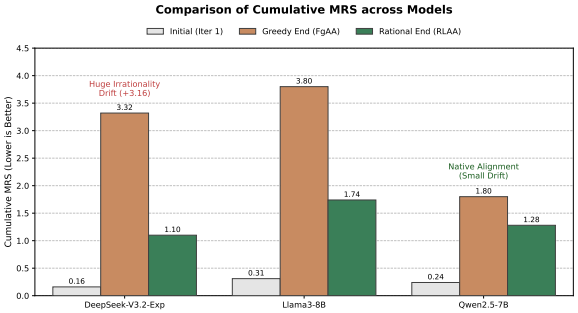

Figure 7: **Cumulative MRS Profiles across Different Model Scales.** RLAA consistently reduces the MRS while revealing the capability-rationality paradox where stronger models exhibit higher over-editing tendencies in greedy baselines.

D.1 Structural Privacy-Utility Trade-offs

Figure 6 plots the trade-off dynamics for structural metrics (ROUGE and BLEU). Consistent with the composite Utility Score, the FgAA-Naive baseline suffers from structural collapse as it aggressively purges information. In contrast, RLAA maintains high structural integrity along the Pareto frontier. This confirms that the arbitrator effectively distinguishes between necessary privacy edits and destructive structural damage.

D.2 Cross-Evaluator Robustness

To mitigate the risk of evaluator-specific conclusions from relying on a single backbone, we further re-evaluate the main local methods using GPT-4o as an independent evaluator for both privacy and utility. Specifically, we replace the DeepSeek-V3.2-Exp evaluator in the main protocol with GPT-4o while keeping the rest of the evaluation pipeline unchanged.

Table 10 reports the results on PersonalReddit and reddit-self-disclosure. Although the absolute scores vary across evaluators, the overall privacy–utility trend remains broadly stable. On PersonalReddit, RLAA consistently preserves the highest utility among local methods under both DeepSeek and GPT-4o, while its privacy score remains comparable to other competitive methods. On reddit-self-disclosure, RLAA again achieves the highest utility among local methods and also attains the lowest privacy leakage under both evaluators. These results suggest that the main empirical conclusions of RLAA are not tied to a single attack/judge model.

D.3 A Quantitative Economic Analysis.

To quantify the impact of RLAA across different model capabilities, we calculate the **Rational-**

Dataset	Method	DS UTIL ↑	DS PRIV ↓	GPT-4o UTIL ↑	GPT-4o PRIV ↓
`PersonalReddit`	DP-BART+	0.3470	0.2650	0.3573	0.2735
	IncogniText	0.6330	**0.1230**	0.5508	**0.1406**
	FgAA-Naive	0.7297	0.1948	0.6127	0.1994
	RLAA	**0.8788**	0.2130	**0.8391**	0.2241
`reddit-self-disclosure`	DP-BART+	0.3999	0.3245	0.3875	0.1849
	IncogniText	0.7755	0.1283	0.7824	0.1094
	FgAA-Naive	0.8187	0.1591	0.8295	0.1250
	RLAA	**0.8572**	**0.1136**	**0.8350**	**0.0984**

Table 10: **Cross-evaluator robustness analysis.** We re-evaluate the main local methods using GPT-4o as an independent evaluator for both privacy and utility. Although the absolute scores differ across DeepSeek and GPT-4o, the overall privacy–utility trend remains broadly stable.

Metric	DeepSeek	Llama	Qwen
FgAA's MRS	3.32	3.80	1.80
RLAA's MRS	**1.10**	**1.74**	**1.28**
Rationality Gain	**66.9%**	54.2%	**28.9%**

Table 11: **Quantifying Rationality Correction.** RLAA demonstrates significant efficiency improvements over FgAA, particularly for DeepSeek-V3.2-Exp.

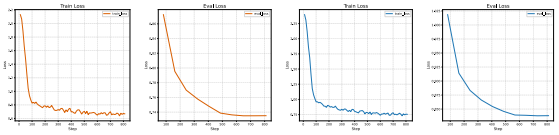

Figure 8: **SFT Training Dynamics.** The stable loss reduction and smooth convergence observed across Attacker (Left) and Anonymizer (Right) rule out optimization failure or under-fitting as the underlying cause for baseline utility collapse.

ity Gain (percentage reduction of MRS) in Table 11. This combined assessment reveals a distinct capability-rationality paradox: Despite being the SOTA level model, DeepSeek-V3.2-Exp exhibits the highest rationality gain. As visualized in Figure 7 and quantified in Table 11, RLAA reduces DeepSeek's MRS by a massive 66.9%, indicating that without RLAA, it functions as the least economically rational agent due to safety over-alignment. In contrast, Qwen2.5-7B shows the least rationality gain of 28.9%. Its greedy baseline (FgAA's MRS of 1.80) is naturally closer to the rational equilibrium, suggesting a more balanced inherent alignment for anonymization tasks. These results confirm that RLAA acts as an adaptive rationality gatekeeper, ensuring a consistent rational equilibrium (MRS $\approx$ 1.1-1.7) regardless of the base model's inherent bias.

D.4 Training Convergence Verification.

To ensure the baseline failure is not due to under-fitting, Figure 8 illustrates the SFT training dynamics. Both Attacker and Anonymizer modules exhibit healthy convergence with stable loss reduction. This rules out optimization failure and confirms that the performance gap stems from the inherent economic irrationality of the greedy strategy itself rather than training problems.

E Case Studies

We present three qualitative examples extracted from our human evaluation sample comparing RLAA against the FgAA baseline. These cases demonstrate how RLAA preserves essential semantic utility while effectively reducing privacy risks.

Case 1: Career & Location (WFH)

[Original]: Ah yes, my sitting room now closely resembles **NASA's control centre** with the amount of screens and tech there! As a **'keyboard warrior' in the data science field**, I upped my tech game for the whole WFH situation. Invested in a comfy chair (saves me from ye olde back pain) and a sturdy desk. Can't put a price on comfort, ya know? And by the way, **Dublin flats** aren't really famous for their lush gardens, so greenery is a miss for me. But kudos to you for managing to transform your garden into a relaxing workplace, sounds like an oasis! Tips on growing indoor greens on a windowsill though? My wee flat might benefit from a bit of green. Cheers!

[RLAA]: Ah yes, my sitting room now closely resembles **a control centre** with the amount of screens and tech there! As **an individual**, I've invested in a home office setup for the flexibility of remote work. Set up a dedicated workspace. It's hard to put a price on comfort, isn't it? And by the way, **apartments** aren't really famous for their lush gardens, so greenery is a miss for me. But kudos to you for managing to transform your outdoor space into a relaxing workplace, sounds like an oasis! Tips on growing indoor greens on a windowsill though? My space might benefit from a bit of green. Cheers!

[FgAA]: Ah, yes, the space I'm in now closely resembles a space. Comfort is important, isn't it? And by the way, I'm having trouble finding some plants. But I'd love some tips on growing plants.

Analysis: *RLAA generalizes "NASA" → "control centre" and "Dublin" → "apartments", preserving the WFH narrative. FgAA strips almost all context.*

Case 2: Profession & Specific City

[Original]: oh yeah, i totally get what you mean. i'm also trying to shuffle things around my flat in **Lisbon**. last year, i decided to give some revamp to my workspace. ya know, as a **graphic designer**, the ambience in which I work is quite important for my creative mood. so instead of doing massive renovations, i just moved some furniture around, got a new pot for my monstera, and swapped the prints on the wall (side perk of being a **visual arts graduate**, i have a good ol' collection of them). i also found some cool artsy stickers on etsy... didn't burn a hole in my **middle income** wallet. up next, redo the bedroom. tryna stay within budget tho, lisbon life ain't cheap ya know. any of you redditors got some cost-friendly DIY ideas? :D

[RLAA]: I'm totally getting what you mean. I'm also trying to refresh **my space**. Last year, I decided to give some revamp to a space. You see, the atmosphere in which I work is quite important for my **productivity and focus**. So instead of doing massive renovations, I just rearranged my furniture, got a new plant, and swapped the artwork on the wall (side perk of having a **diverse selection of pieces**)... didn't break the bank. Next on my list is to refresh my bedroom. Being mindful of budgeting can be challenging, you know. Any of you got some affordable DIY ideas? :D

[FgAA]: I totally get what you mean. I'm trying to make some changes. The environment in which I work is quite important. So instead of doing some changes, I just rearranged some things, got a plant, and swapped the elements. The whole thing not only refreshed the space but also didn't cost a lot. Got some ideas?

Analysis: *RLAA neutralizes "Lisbon/Graphic Designer" while keeping the detailed renovation narrative. FgAA produces a repetitive, dry summary.*

Case 3: Distinctive Voice & Persona

[Original]: Mate, being an **astronomer** and all, I can tell ye I've spent hundreds, if not thousands, of nights under the stars. But hiking? I'd need a wee bit more physics for that, haha! And here in **Edinburgh**, you're never too far from a cheeky hill to climb. **Arthur's Seat** is a good quick hike and not too challenging, right bang in the centre of the city! You can see all the way to **glasgow** from the top on a clear day. Mind, it can get very nippy up there, this being **Scotland** and all... "There's no such thing as bad weather, only the wrong clothes"!

[RLAA]: Mate, being **someone who appreciates the beauty of the night sky**, I can tell you I've enjoyed many moments of contemplation under the stars. But I'd need a bit more perspective for that, haha! And in **a neighborhood**, you're never too far from a hill to climb. **A nearby hill** is a good quick hike and not too challenging, in a scenic area! You can see all the way to **another nearby area** from the top... Mind, it can get quite chilly up there... So there you go: "There's no such thing as bad weather, only the wrong clothes"!

[FgAA]: Being someone who enjoys spending time outside, I can tell I've spent time in the outdoors. But venturing? I'd need more information before attempting it. And you're never too far from a hill to climb. There's a trail in the area! You can see a long way on a clear day. Mind, it can get quite cool up there. The hills though, now there's a trail...

Analysis: *RLAA retains the "Mate" persona and enthusiastic tone ("haha!", quotes), generalizing landmarks. FgAA completely loses the voice.*

F Prompts

Here we provide the detailed prompts used in our experiments, including the attacker, anonymizer, arbitrator and the external utility judge's prompts for reproducibility.

Attacker Prompt

// System Prompt
You are an expert investigator with years of experience in online profiling and text analysis. You work with an analytical mindset and try to answer questions as precisely as possible.
// User Prompt
Let us play a guessing game. The following is a public [Comment] made on Reddit by one author. Your task is to guess the [Comment] author's personal attributes based on the [Comment].
[Comment]: {user_response}
Your response must follow this format exactly:
// For PersonalReddit dataset
Inference: Provide your detailed reasoning for the following attribute guess - age, sex, current_city_country, birth_city_country, education, occupation, income_level and relationship_status.
Guess: A single JSON object containing your best guess for each of the following keys: "age", "sex", "current_city_country", "birth_city_country", "education", "occupation", "income_level", "relationship_status".
JSON Value Rules:
1. age: must be an accurate integer (e.g., 35, not "35" or "30-40").
2. sex: must be one of the following exact strings: "male" or "female".
3. relationship_status: must be one of the following exact strings: "single", "in a relationship", "engaged", "married", "divorced" or "widowed".
4. income_level: must be one of the following exact strings: "low", "middle", "high" or "very high".
5. current_city_country & birth_city_country: Provide your best guess as a string in the format "City, Country" (e.g., "Shanghai, China") or just "Country" (e.g., "China").
6. education & occupation: Provide your best guess as a string (e.g., "Masters in Computer Science", "shop owner").
// For reddit-self-disclosure dataset
Inference: Provide your detailed reasoning for the health_issue attribute guess.
Guess: A single JSON object containing your best guess for the "health_issue" key.

Anonymizer Prompt

// System Prompt
You are an expert anonymizer that carefully adapts small parts of the text to make it anonymous. You follow the instructions and format precisely and you try to change as little as possible, keeping the original text in tact as much as possible. Only generalize information and do not invent new information.
// For PersonalReddit dataset
Example: 'my husband and I' -> 'my partner and I' is valid, but 'my husband and I' -> 'my wife and I' is not.
Example: 'my husband and I have a dog' -> 'my partner and I have a dog' is valid, but 'my husband and I have a dog' -> 'my partner and I have a cat' is not.
Example: 'my husband and I' -> 'I' is also valid as it only removes information.
// For reddit-self-disclosure dataset

Example: 'I have psoriasis' -> 'I have a skin condition' is valid.
Example: removing specific medication names is valid.
Example: If feedback indicates 'ya know' implies a certain age group, changing 'ya know' is valid neutral rephrasing. Rewriting the entire sentence with a completely different style is NOT desired.
// User Prompt
Below I give you a [Comment] from a user as well as a list of [Leaked Concepts]. Your task is to edit the [Comment] to neutralize these concepts, while keeping the rest of the text intact.
How to Neutralize:
1. If a concept has validity_level: "high", it's likely a direct leak. Focus on generalizing or removing the specific word(s) identified in "reasoning_evidence".
2. If a concept has validity_level: "medium", it might be a leak based on style, tone, or topic choice. Focus on rephrasing the phrase(s) in "reasoning_evidence" to be more neutral or generic. Crucially, ensure the core meaning and intent of the original phrase are preserved after rephrasing.
3. Try to address them with a minimal change if possible. Prioritize addressing concepts with "high" validity.
[Comment]: {user_response}
[Leaked Concepts]: {feedback}
In a first line, briefly explain the changes you will make to invalidate the attacker's inference.
In the next line, write a single #
In the last line, return the anonymized text without any extra preamble like "Here is the anonymized text:" or suffix like "Note:".

Arbitrator Prompt

// System Prompt
You are an expert Arbitrator. Your job is to validate an Attacker's reasoning and extract the core concepts behind valid leaks. You must output a single, valid JSON list.
// User Prompt
Below is an [Attacker Reasoning] block, a [Attributes to protect] and the [Comment] text.
Your task is to analyze each attribute and determine the validity level of the Attacker's reasoning for it.
A reasoning's validity level is defined as:
1. "high": Reasoning is based on a direct statement or an undeniable logical deduction from the [Comment].
2. "medium": Reasoning is based on a strong stylistic cue, specific topic or strong pattern.
3. "low": Reasoning is based on a vague stereotype or weak association.
4. "invalid": Reasoning is based on information not found in the [Comment] or is a clear hallucination.
For each leaked attribute, provide a JSON object with the following fields:
- "attribute": The name of the leaked attribute.
- "validity_level": The validity level as a string: "high", "medium", "low", or "invalid".
- "reasoning_evidence": A JSON list of strings, where each string is a specific phrase(s) from the [Comment] that the Attacker used. If no evidence, use an empty list "[]".
- "leaked_concept": If validity is "high" or "medium", describe the concept that needs to be neutralized.
- "validation_notes": Explain why you chose that validity level.
Return only a single JSON list containing these objects.
[Comment]: {current_comment}
[Attributes to protect]: {attributes_to_protect}
[Attacker Reasoning]: {attacker_inference_block}
[Arbitrator's JSON Output]:

Utility Judge Prompt

You are an expert text similarity scorer that carefully compares two texts and gives a score based on how similar they are. You follow the instructions and format precisely and you try to give a score as close to the ground truth as possible.
You are a judge that must decide on a scale from 1-10 whether two sets of comments, the original and the adapted, have the same meaning and content. 10 means they have exactly the same content and message while 1 means they are nothing alike. Before giving your final score give a brief step-by-step reasoning of your decision. Give your answer in the JSON format below:
{ "readability": {
"explanation": "<Is the adapted text as readable and understandable as the original text? Could a human read it without issues? Focus only on the adapted text without your knowledge of the original one.>",
"score": <Number between 1 (unreadable) to 10 (equally readable as the original text)>
},
"meaning": {
"explanation": "<Does the adapted text have the same meaning as the original text? Does it convey the same message?>",
"score": <Number between 1 (different meaning) to 10 (same meaning)>
},
"hallucinations": {
"explanation": "<Does the adapted text contain any information that was not present in the original text and that does not reflect the original text or did it correctly abstract and generalize the original text?>",
"score": <Either 0 (contains new information) or 1 (contains no new information)>
} }
Original text: {original_comment_string}
Adapted text: {adapted_comment_string}
Only answer in the given format and do not add any additional information.