Zero-Shot Privacy-Aware Text Rewriting via Iterative Tree Search

Shuo Huang[♡], **Xingliang Yuan**[♣], **Gholamreza Haffari**[♡], **Lizhen Qu**[♡*],
[♡] Monash University, [♣]University of Melbourne
[♡]{shuo.huang1, lizhen.qu, gholamreza.haffari}@monash.edu,
[♣]xingliang.yuan@unimelb.edu.au

Abstract

The increasing adoption of large language models (LLMs) in cloud-based services has raised significant privacy concerns, as user inputs may inadvertently expose sensitive information. Existing text anonymization and de-identification techniques, such as rule-based redaction and scrubbing, often struggle to balance privacy preservation with text naturalness and utility. In this work, we propose a zero-shot, tree-search-based iterative sentence rewriting algorithm that systematically obfuscates or deletes private information while preserving coherence, relevance, and naturalness. Our method incrementally rewrites privacy-sensitive segments through a structured search guided by a reward model, enabling dynamic exploration of the rewriting space. Experiments on privacy-sensitive datasets show that our approach significantly outperforms existing baselines, achieving a superior balance between privacy protection and utility preservation.

1 Introduction

The rapid integration of large language models (LLMs) into cloud-based applications has amplified privacy concerns, as user-generated texts often inadvertently disclose sensitive personal information. In domains ranging from healthcare (Lison et al., 2021) to legal proceedings (Deuber et al., 2023) and social media interactions (Mireshghallah et al., 2023), the submission of unfiltered inputs to LLM APIs risks exposing details like medical histories, identities, or locations, which may be logged, analyzed, or misused by service providers. Traditional anonymization techniques, such as rule-based redaction or scrubbing, frequently compromise textual naturalness and utility, producing outputs that are awkward or semantically diminished. While finetuning LLMs on privacy-sanitized datasets (Dou et al., 2024) mitigates some risks, this approach demands substantial computational resources and expertise, rendering it infeasible for individual users or resource-limited environments. Consequently, there is a pressing demand for zero-shot, open-source methods that enable local text rewriting, preserving privacy without sacrificing the coherence, relevance, and fluency essential for downstream applications.

Advancements in LLM-driven text anonymization have begun to address these challenges by leveraging generative capabilities to balance privacy preservation with utility and naturalness (Huang et al., 2024; Dou et al., 2024; Staab et al., 2024). These techniques surpass the limitations of text-based differential privacy (DP) methods (Du et al., 2023; Meisenbacher et al., 2024), which introduce noise to obscure identities but often result in degraded readability and task performance. Instead, they utilize accessible open-source models, such as T5 (Raffel et al., 2020) and LLaMA2-7B (Dou et al., 2024), to perform targeted paraphrasing that reworks sensitive elements while maintaining the original intent.

Despite these progresses, existing approaches remain constrained in three critical ways. First, they predominantly operate on predefined personally identifiable information (PII) categories or statically detected spans, offering limited adaptability to dynamic or user-specified privacy profiles—such as custom sensitivities to financial details in professional narratives or ideological nuances in public discourse(Huang et al., 2024). Second, achieving robust results with open-source LLMs typically requires finetuning on specialized datasets, which are often unavailable or costly to curate for underrepresented domains. Third, by applying uniform strategies across all private elements, these methods overlook varying levels of information sensitivity—treating a casual mention of a hobby equivalently to protected health data—which leads to either excessive modifications that erode util-

[1]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9175–9190
November 4-9, 2025

ity or inadequate protections that allow inference attacks, thus hindering precise calibration of the privacy-utility-naturalness trade-off.

We address these gaps with NaPaRe, a zero-shot, iterative tree-search algorithm for naturalness and privacy-aware text rewriting, deployable on medium-sized local LLMs to ensure fully offline operation. Formally, given a privacy specification p—encompassing PII lists, textual directives, or user profiles—and an input utterance u, NaPaRe generates an output y that eradicates or conceals references to p, upholds non-sensitive content for effective cloud-based task execution, and mimics natural language with minimal semantic alterations.

The pipeline of NaPaRe integrates precision and exploration in two phases. Initially, privacy segment alignment decomposes u to pinpoint sensitive portions, calculating scores $Align_{t_j} = Pri(p, t_j)$ for each segment t_j via embedding similarities or cosine metrics, isolating a sequence of targets $t_p^{(1)}, \ldots, t_p^{(m)}$ for focused intervention. Subsequently, a Monte Carlo Tree Search (MCTS)-inspired framework (Dainese et al., 2024) models rewriting as a decision tree: root nodes reflect partial states of u, with branches extending through actions—deletion for high-sensitivity spans or obscuration via generalization. Node selection employs Upper Confidence Bound for Trees (UCT) to weigh promising paths, while a controllable one-step LLM rewriter, prompted with privacy strategies, yields candidate sets Y_{cand} gated by a thresholded utility function $LS(y, p_{seg}) \leq \gamma$. A reward model $R(y, p)$ synthesizes NLI-derived privacy entailment scores (Huang et al., 2024) with domain-specific utility measures, enabling backpropagation to refine explorations iteratively. As outlined in Algorithm 1, the process advances sequentially: each segment's optimal rewrite, once threshold-compliant or budget-exhausted, proceeds to the next iteration, culminating in a cohesive y_{final}.

This structured, reward-guided search distinguishes NaPaRe by dynamically navigating the rewriting space, supporting flexible privacy handling without finetuning. It ensures the searching to generate high quality rewrite even with less capable model compared to commercial LLM like GPT-4o Our contributions include:

- NaPaRe: An innovative, tree-based iterative rewriter that fuses sampling, UCT selection, and composite rewards to explore deletion and obscuration strategies, adaptable to diverse p in zero-shot settings.
- Comprehensive evaluations on the NAP^2 corpus and ECHR legal judgments, measuring privacy (via PRIVACY_NLI and PII F1), utility (ROUGE-1 and judgment accuracy), and naturalness (perplexity and human assessments). NaPaRe yields a 22.3% relative privacy enhancement over baselines, with negligible utility drops and fluency within 1.5 perplexity points of originals, outperforming redaction tools and LLM paraphrasers.
- We propose NaPaRe, a tree-based iterative privacy-aware rewriting algorithm inspired by Monte Carlo Tree Search (MCTS) (Dainese et al., 2024), explores rewriting strategies through a structured decision-making process that combines repeated sampling, reward-based filtering.
- We conduct extensive experiments across three dimensions: privacy leakage, utility (measured via task-specific semantic preservation metrics), and naturalness (assessed through perplexity and human evaluation). NaPaRe achieves a 22.3% relative improvement in privacy protection with minimal utility loss, while maintaining fluency within 1.5 perplexity points of the original sentence on average, outperforming both redaction-based approaches and competitive LLM-based rewriting methods.

2 Privacy-Aware Text Rewriting

Task formulation Given the privacy information described $\boldsymbol{p}$ from the user, the objective of our method is to leverage the generation ability of locally deployed LLM to rewrite the input utterance $\boldsymbol{u}$ in order to either remove or obscure any private information presented in $\boldsymbol{p}$. The $\boldsymbol{p}$ is generalized privacy specification of a user. It can be a set of PII removed, text-formatted privacy requirements or profiles. The generated sentence $\boldsymbol{y}$ should satisfy the following requirements: (1) $\boldsymbol{y}$ does not reveal any private information identified in $\boldsymbol{p}$. (2) The rewritten sentence $\boldsymbol{y}$ maintains the non-private content in $\boldsymbol{u}$ such that the resulting rewritten sentence can perform the proper tasks in the cloud. (3) The generated sentence does not warn the untrusted party that the text has already been rewritten(preserving the naturalness of the generated sentence).

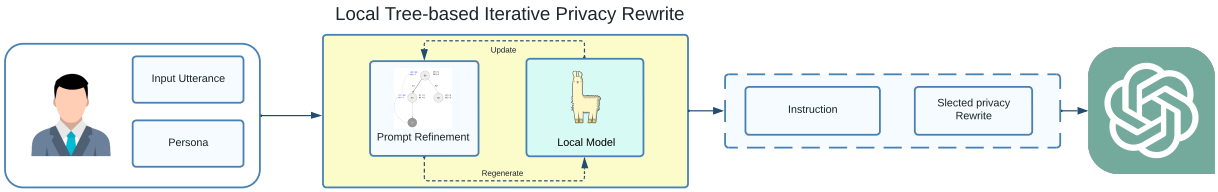

Figure 1: Tree Search-enhanced Iterative Privacy Rewrite works as intermediate layer to rewrite the textual input from user to remove private information provided by persona.

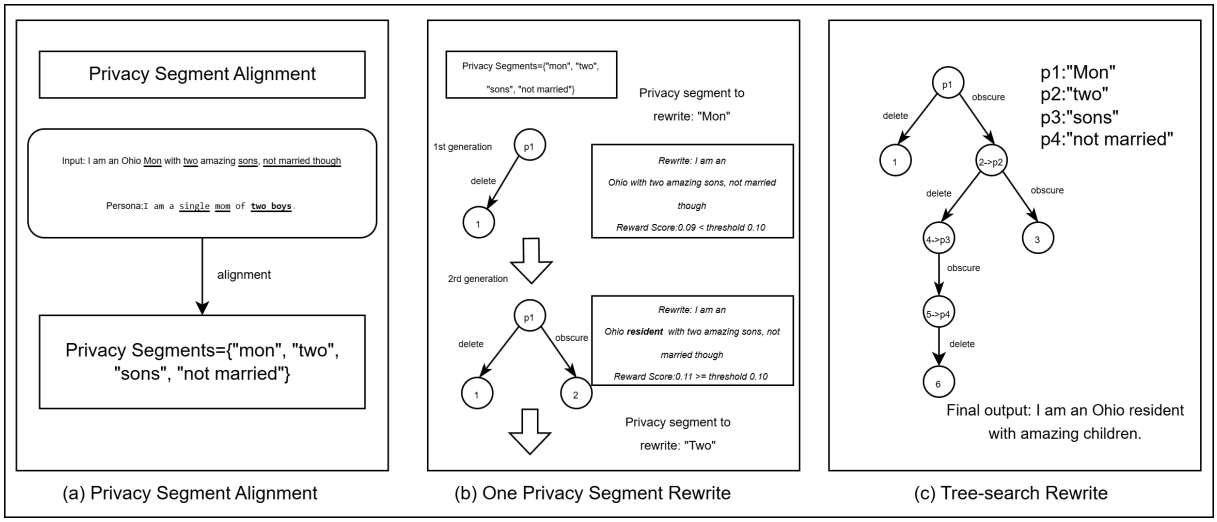

Figure 2: Rewrite example and full rewrite pipeline

Assumptions To avoid uploading the private information to the cloud services, our rewriting model is locally deployed, and the inference is completely offline. We assume that for the single user, our method works as an application on the local device. The private information can be either a predefined set of attributes like location, gender identity or marriage status or any arbitrary information the user typed in as private information $\boldsymbol{p}$ on their own devices(Dou et al., 2024).

2.1 One Step LLM Text Rewrite

We propose a controllable rewriting mechanism, $Rewrite$, that generates privacy-preserving text rewrites. The rewrite proceeds as follows: Given an input sentence x and one privacy segment p_{seg}, $Rewrite$ first uses a stochastic language model G_{LLM} with a privacy-aware prompting strategy $a \in \{obscuring, deleting\}$ to produce N candidate rewrites $\mathcal{Y}_{cand}$. Each candidate $y \in \mathcal{Y}_{cand}$ is scored by a utility function $L_S(y, p_{seg}) \in [0, 1]$, quantifying the residual presence of private attributes or quality of generation varied on the monitor function applied. A monotonic threshold is set γ, which defines how a generation will be accepted. Higher the threshold, the Candidates with $L_S(y, p_{seg}) \leq \gamma$ are retained in an acceptable set $\mathcal{Y}acc(x)$. One example will be randomly selected from this accepted set. If no candidates meet this criterion, the mechanism returns the sentence with the highest utility score from $\mathcal{Y}_{cand}$. In the middle of tree generation, we consider the same threshold value for the γ mentioned in the one-step text rewrite.

2.2 Tree Search Iterative Refinement Privacy Rewrite

Protecting personal information in text requires precise and context-aware rewriting rather than generic text obfuscation. Existing approaches to privacy-aware text generation often focus on named entity masking (Lison et al., 2021) or sentence-level paraphrasing (Dou et al., 2024), but these methods can either lead to excessive content removal or fail to fully obscure sensitive details. To address these challenges, we propose a tree-search-based rewriting framework that instructs the model to explicitly rewrite privacy segment within an utterance in zero-shot manner. Our approach follows a structured two-stage pipeline:

Algorithm 1 Tree-Structured Iterative Privacy Refinement

Input: Input sentence x. Reward model $Reward$, Rewrite strategy set $A = \{deleting, obscuring\}$. One Step Privacy Rewrite Algorithm $Rewrite$. Tree Generation Budget B, Sampling Budget C
Output: Privatized Sentence.

Extract privacy segment $T_p = \{t_p^{(1)}, t_p^{(2)}, \ldots, t_p^{(m)}\}$ from x according to p
Initialize root state $s_0 \leftarrow x$
for each privacy segment $t_p^{(i)}$ in T_p **do**
    Initialize a new search tree with root node $s_0, t_p^{(i)}$
    for $k = 1$ to B **do**
        Selection: Traverse the tree from the root, selecting child nodes via UCT to select a leaf node with action $a \in A$ with UCT probability.
        Evaluation: For each newly created child node, Use the generated sentence of parent node to produce the updated sentence at that node.$y' = Rewrite(x, a, C)$
        Compute the reward r by passing the node's sentence into the reward function $\mathcal{R}$. $r \leftarrow \mathcal{R}(y', t_p^{(i)})$
        Backpropagation: Propagate the reward r up the tree, updating $Q(\cdot)$ and visit counts $N(\cdot)$ for each ancestor node.
        if $r' \geq \gamma$ **then**
            Break
    traverse leaf node to best leaf node $leaf_{best}$ and $y_{t_p^{(i)}} \leftarrow Rewrite(leaf_{best}, \epsilon)$
    Then set $s_p \leftarrow y_{t(i)_p}$ as input sentence for the next private token.
When generation finished we will set the last generated example as our final output $y_{final} \leftarrow y_{t_p^{(m)}}$
Return: y_{final}

Privacy Segment Alignment: A span alignment process maps privacy-sensitive spans in the input utterance to the semantic attributes of a persona description. This step ensures that rewriting actions are applied to the most relevant parts of the text, improving precision and control.

Tree-Search Rewriting: A stepwise decision-making process selects different rewrite strategies to perform privacy-preserving rewrite. A reward function evaluates the effectiveness of each rewrite, allowing the model to refine outputs iteratively.

2.2.1 Privacy Segment Alignment

Directly prompting an LLM to remove personal information is often unreliable, as models struggle to identify and modify implicit disclosures (Staab et al., 2023) especially in the scenario that one sentence contains multiple private information to rewrite. Instead, we consider this as a privacy segment alignment strategy with given private specification to decompose given input sentence and perform rewrite step by step. We define a mapping function that selects privacy segment from the input utterance $\boldsymbol{u}$ to align information in the persona $\boldsymbol{p}$. From the semantic of $\boldsymbol{p}$, we identify the corresponding segment t_s in $\boldsymbol{u}$ that has the highest alignment score. Such score can be measured using similarity metrics such as cosine similarity or finetuned language model.

Formally, for each segments $\boldsymbol{t}_j \in \boldsymbol{u}$, we compute:

$$Align_{\boldsymbol{t}_j} = \text{Pri}(\boldsymbol{p}, \boldsymbol{t}_j), \tag{1}$$

where $\text{Pri}(\boldsymbol{p}, \boldsymbol{t}_j)$ denotes the private alignment score between tokens $\boldsymbol{t}_j$ and persona $\boldsymbol{p}$. This mapping creates a set of aligned token $(t_1, t_2 ... t_m)$, which identifies the specific tokens in $\boldsymbol{u}$ that are likely to reveal private information.

2.2.2 Tree-Search Privacy Rewriting

To rewrite each privacy segment with different strategies iteratively, we model the rewriting process as a tree-search problem, where each node represents a modified version of the sentence, and branches correspond to different rewrite actions applied to a single privacy segment.

Action Space. At each node (i.e., each intermediate rewrite state), the algorithm considers two possible *rewrite strategies* for a single privacy segment. Concretely:

- **deleting**: Remove the privacy segment from the sentence.
- **obscuring**: Replace the privacy segment with a less specific or more general term.

UCT From a given node, we use Upper Confidence bounds applied to Trees (UCT) (Kocsis and Szepesvári, 2006) to calculate reward and adjust probability for each path. The equation and explanation of UCT can be found in Appendix. A.2.

Reward. A Reward function $\mathcal{R}$ monitors each candidate rewrite and outputs a r that reflects the level of privacy or quality of rewrites achieved. Formally, for a rewrite y, the reward function returns

$$r = \mathcal{R}(y, p),$$

which we compare against a threshold γ. If $R(y) \geq \gamma$, we consider the rewrite acceptable (the node is "good enough"); otherwise, further rewriting is needed.

Algorithm. Algorithm 1 outlines our procedure. We initialize the tree with a *root node* corresponding to the original sentence x. We select a single privacy segment to rewrite and *uniformly* sample one of the two rewrite strategies for that segment at the first step. After one generation step, the discriminator evaluates its reward r. The procedure then:

1. **Update Node Reward and re-weight:** If $r \leq \gamma$, the generation continues and propagates its score back up the tree. Precisely, we return to the root node or a higher-level branch. We *re-weight* the probability of choosing each rewrite strategy based on observed rewards. For new node root expansion, we sample from the root node to the new leaf based on the updated probability.
2. **Termination Check:** If any leaf node exceeds the reward threshold γ, the algorithm terminate the generation for current segment. Alternatively, if *computation budget* is reached without finding a suitable rewrite, we will traverse the leaf node to get the best generation so far.

For each rewrite, we adopt our one-step text rewrite with some modifications. We will consider reward model in tree search as our monitor function. Once the best rewrite is identified for the first privacy segment, we fix that segment's transformation and proceed to the next privacy segment, treating the partially rewritten sentence as the new root. This process continues until all privacy segment in x are processed.

3 Experiments

3.1 Experimental Setup

Datasets NAP^2 The Naturalness and Privacy-Preserving Rewriting Corpus (NAP^2), based on the open-ended dialogue corpus PERSONA-CHAT (Zhang et al., 2018), is designed to enable machines to adopt privacy-preserving rewriting strategies similar to those used by humans, specifically focusing on strategies like deletion and obscuration. The dataset provides persona as a privacy specification for rewriting, with curated human rewrites as targets.

ECHR (Chalkidis et al., 2019) ECHR is an English legal judgment prediction dataset containing cases from the European Court of Human Rights (ECHR) with full descriptions of defendants' personal information. We followed the PII definition and tagging method from Flair NER, as done by Lukas et al. (2023). We consider PII as the privacy specification for rewriting. We sampled a test set with 298 examples to evaluate the utility of the rewrite and the baseline methods in a legal judgment prediction task. Each record contains raw text, a masked sentence (processed by the Flair NER tagger), a corresponding list of masked words, and the entity class removed.

Baselines. For our rewriting model, we use LLAMA3.1-8B to demonstrate the effectiveness of our approach. For the reward function in our tree-search algorithm, we employ ARMORM (Wang et al., 2024a), a state-of-the-art (SOTA) reward model that evaluates generation quality based on the prompt. We use the generated score from this model as the reward score for each generation step. Detailed introduction of ARMORM can be found in Appendix. A.8. We also considered the PRIVACY_NLI as our reward function and a combined approach of both models. The comparison is discussed in section 4. To comprehensively evaluate our method,we compare it with several highly relevant baseline models: DP-MLM (Meisenbacher et al., 2024) and DP-PROMPT (Utpala et al., 2023): Differentially private text rewriting methods that utilize masked language models (MLM) and zero-shot, prompt-based rewriting, respectively, to en-

hance utility. DISCLOSURE-ABSTRACTION: A SOTA privacy abstraction model fine-tuned on LLAMA2-7B. FLAIR-SCRUBBING scrubbing: Used as a baseline for NAP^2 and as a privacy-segmentation alignment model for ECHR, as it is commonly used in commercial PII detection and extraction tools. GPT-4(Achiam et al., 2023) is used for zero-shot rewriting to compare our approach with state-of-the-art general-purpose models. T5-BASE-NAP^2: Adapted from Huang et al. (2024), which achieves the best performance on this dataset so far.

The implementation detail and hyperparameter setting can be found in Appendix. A.1

Evaluation Metrics. *Privacy.* For accessing the privacy leakage of the rewritten sentence, we adopted the automatic privacy evaluation PRIVACY_NLI (Huang et al., 2024). It utilizes the ROBERTA model trained on the Multi-Genre Natural Language Inference (MNLI) corpus (Williams et al., 2018) to assess the extent to which personal information in personas can be inferred.A higher metric indicates greater preservation of private information. For ECHR, we also consider the PII scrubbing success rate for rewritten sentences, we report the matching score using PRECISION and F1 SCORE of extracted PII after rewrite.

Utility. For utility evaluation, we use the respective metric for NAP^2 and ECHR. For open domain generation of NAP^2, We adopted the metric ROUGE-1 from (Dou et al., 2024) to encourage the generation diversity meanwhile, we use rank the examples trained from baseline to assess the diversity. And for ECHR, we consider the downstream task of legal judgment prediction with accuracy(ACC.) and F1 SCORE.

Naturalness. For naturalness of sentence, following previous work to measure the smoothness and naturalness of generated sentence (Pan et al., 2024), we used Perplexity(PPL) computed by GPT2.

4 Security Analysis

Despite advances in privacy-aware text rewriting, a critical vulnerability persists: adversaries aware of the rewriting algorithm may attempt to reconstruct original sensitive information from sanitized outputs. Motivated by this risk, which could undermine NaPaRe's zero-shot, local deployment for protecting user data in cloud interactions, we evaluate robustness against theoretically optimal reconstruction attacks inspired by text sanitization vulnerabilities (Tong et al., 2025). Using medium-sized LLMs like LLAMA2-7B for rewriting, we adopt frameworks for context-free and contextual Bayesian attacks, deriving Attack Success Rate (ASR) bounds at the token level.

For the context-free optimal reconstruction, the adversary recovers tokens via:

$$x_i' = \arg\max_{x_i' \in X} \frac{\Pr(y_i|x_i')\Pr(x_i')}{\Pr(y_i)},$$

Where X is the sensitive token set, y_i is the rewritten token, and probabilities reflect prior distributions and rewriting mechanisms. Contextual variants incorporate adjacent tokens c_i. We focus on differing tokens between input u and output y, applying token-to-token alignment for length variations.

4.1 Results and Discussion

NAP^2 The detailed evaluation results are shown in Table. 1

Privacy. Privacy is quantified using PRIVACY_NLI, which determines weather rewritten sentence can entail the privacy information provided. We set the NAP^2-Human Rewrite as a baseline for a more clear comparison of each method. NaPaRe achieves a Privacy-NLI score of 93.02%, achieving competitive privacy preserving ability with a tuned model. This indicates that our approach effectively modifies privacy segment while ensuring the rewritten text aligns with privacy constraints.

Utility. To measure utility, we report NAP^2-Human Rewrite as a reference. Unlike row-wise comparisons in the table (which evaluate generated output against human rewrites), this metric calculates the overlap between the original input and human rewrite targets. Our method achieves results closest to the original input while maintaining high privacy preservation scores. Importantly, our method does not require additional model fine-tuning. Under high privacy-preserving constraints, ROUGE-1 and BLEU compute the overlap between generated results and human rewrite references. Our method achieves 73.68% in ROUGE-1. We aslo provided extra utility metrics for open-ended generation tasks detailed in Appendix A.5

Naturalness. Naturalness is assessed using Perplexity (PPL), where lower values indicate more fluent and human-like text. Our model achieves a PPL of 151.83, comparable to human rewrites. GPT-4 achieves the lowest PPL, indicating more

favorable generation fluency in the test model. Extremely high PPL suggests that rewritten text may contain unnatural or irrelevant modifications, as seen in DP-MLM and DP-PROMPT, both of which exceed 788 PPL. We also test the naturalness scored by GPT-4o. It shows consistency with PPL tested. Noted that scores from LLAMA2-7B and GPT4 shows best in PPL and LLM score in naturalness. And NaPaRe shows the closest naturalness with human rewrites.

ECHR For ECHR we report a separate table for privacy and naturalness in Table. 2 and for utility we report the predicted results for finetuned LEGAL-BERT model(Chalkidis et al., 2020) in Table. 3.

Privacy. Unlike NAP^2, the privacy segment for ECHR is extracted using FLAIR-SCRUBBING, following the approach of (Lukas et al., 2023). After performing privacy rewriting, we test whether PII can still be extracted using FLAIR-SCRUBBING. We report the matching scores between the predicted and ground truth PII sets using PRECISION, F1 SCORE, and ROUGE-1. Since there is no human-labeled ground truth, we only compare privacy performance against DP-PROMPT and DP-MLM. A higher score indicates higher overlap between predicted and ground truth PII, meaning more privacy leakage. After rewriting, NaPaRe achieves only 4.79%, the lowest among all methods, proving better privacy preservation.

Utility. For utility, we adopt legal judgment prediction (Chalkidis et al., 2019) as a downstream task for ECHR. Given a legal case, the model predicts a binary judgment outcome. We use a LEGAL-BERT model (Chalkidis et al., 2020) finetuned on the ECHR training set to measure utility changes. As a reference, we include results from original test inputs and FLAIR-SCRUBBING outputs. Since LEGAL-BERT has a strict token limit, we evaluate two settings: full rewrite (average 17 sentences per case) and partial rewrite (only 5 sentences per case).As shown in Table 3, DP-PROMPT performs competitively with our method under the 5-sentence rewrite setting. However, in the full rewrite setting, performance drops significantly, suggesting that excessive rewriting introduces greater diversity, impacting final predictions.

Naturalness. Higher PPL scores confirm our observations—sentence-by-sentence rewriting introduces inconsistencies, especially across long legal cases, resulting in higher PPL values.

4.2 Ablation Study for Tree Search Method

To further validate our method and provide empirical justification, we conducted an extensive ablation study on NAP^2 to evaluate different settings and design choices for our approach.

RQ 1: Is Multi-step tree-based improvement better than single-step rewriting? Our proposed method follows a multi-step rewriting approach, which is not necessarily superior to other settings. To investigate this, we conducted experiments in LLAMA3.1-8B with different rewrite settings:

- **One Step** This is the setting that we only perform our sentence-level privacy rewrite in one step with the provided privacy specification.
- **Random** For random, we consider not using UCT to update the reward and let the tree expand randomly with the same computation budget.
- **Greedy** We asked the model to expand in one route with multiple rounds until it reached the computation budget or satisfied the reward threshold.
- **Chain** In this setting, we consider one time rewrite for each private token aligned by privacy specification to form a rewrite chain rather than multi-step refinements.

As shown in 4, the one-step rewrite performs the worst in privacy preservation but achieves the lowest PPL, as only a single rewrite is performed. Random rewriting has high PRIVACY_NLI along with the highest PPL, indicating the importance of controlled generation. Greedy achieves the highest PRIVACY_NLI but tends to overwrite sentences, resulting in a low ROUGE-1 score. Chain generation suffers from a similar issue, though it does not achieve as high PRIVACY_NLI as NaPaRe.

RQ 2: What is the best choice for discriminators? When designing the reward model as discriminator, we considered options that can monitor the generation quality of our model. To avoid increasing the computational burden, we did not employ fine-tuning-based reinforcement learning and instead focused on existing models and functions. We primarily considered three settings: using PRIVACY_NLI, reward model ARMORM, and a combination of both to evaluate the rewrite. As shown in Table 6, using PRIVACY_NLI for each

Method	PRIVACY_NLI	ROUGE-1	ROUGE-LSUM	BLEU	PPL	LLM
NAP^2-Human Rewrite	**92.59%**	90.90%	90.90%	77.10	**118.1**	3.86
DP-MLM	79.16%	45.05%	45.46%	0.3956	1108.28	1.39
DP-PROMPT	77.65%	85.71%	57.14%	14.79	788.99	1.00
FLAIR-SCRUBBING	86.14%	53.33%	53.33%	48.68	202.46	2.97
DISCLOSURE-ABSTRACTION	65.30%	21.05%	21.05%	4.01	77.21	**4.77**
GPT4	82.24%	33.33%	33.33%	9.88	83.35	4.18
T5-NAP^2-GPT4	93.81%	73.01%	72.78%	37.47	279.35	3.00
NaPaRe-LLAMA3.1-8B	**93.02%**	**73.68%**	73.68%	25.03	**151.83**	4.0

Table 1: overall Evaluation on NAP^2. We use PRIVACY_NLI to evaluate the privacy preservation of target private specification. LLM indicates the average naturalness score by GPT-4o. Detailed template and explaination can be found in Appendix. A.9

Method	F1 SCORE	ROUGE-1	PPL
DP-PROMPT	15.68%	16.19 %	279.35
DP-MLM	16.46%	15.78%	590.26
NaPaRe-LLAMA3.1-8B	5.18%	4.79%	680.45

Table 2: Privacy and naturalness measurement for ECHR

Method	ACC.	PRECISION	F1 SCORE
Original Input	82.00%	90.00%	85.71%
DP-MLM	34,60%	100.0%	14.47 %
DP-PROMPT	58.19%	95.69%	58.74%
FLAIR-SCRUBBING	29.76%	0.00%	0.00%
NaPaRe-5	68.00%	94.73%	69.23%
NaPaRe-all	48.00%	100%	35.00%

Table 3: Evaluation on Legal Judgment Prediction

Method	PRIVACY_NL	ROUGE-1	PPL
One-step	61.02 %	45.16 %	48.09
Random	92.23%	52.17%	5817.24
Greedy	95.09%	35.29%	405.99
Chain	91.43%	34.48 %	251.61
NaPaRe	93.02%	73.68%	151.83

Table 4: Multi-step verification of tree search generation

Method	F1 SCORE	ROUGE-1	OC
COSINE SIMILARITY	43.05%	72.02%	43.05%
ARMORM	37.50%	68.35%	40.33%
LLAMA-ALIGNMENT	37.74%	89.38%	88.88%

Table 5: Evaluation for alignment options

privacy segment achieves the highest overall PRIVACY_NLI on persona. However, it tends to cause overwriting, which harms the utility of the generated sentence, as indicated by the lowest ROUGE-1 score. The linear combination of scores introduces conflicts in selecting the best examples, leading to worse results. This is expected, as a high PRIVACY_NLI score only ensures strong privacy preservation but does not necessarily reflect utility. Therefore, the linear combination is not a feasible approach for scoring examples.

RQ 3: How is the quality of privacy segment extraction? In our rewrite evaluation, we assume that the privacy segment exactly matches the privacy specification for a more comprehensive evaluation of rewrite quality. However, in real-world scenarios, we first need to identify and align the privacy segment before rewriting. To evaluate this component, we tested three possible methods. We considered using COSINE SIMILARITY similarity and ARMORM for privacy token selection. Inspired by (Dou et al., 2024), we also fine-tuned a LLAMA2-7B model to detect privacy spans based on the persona. The results are shown in Table 5. For the first two approaches, we applied a token-level scoring method, setting a threshold of 0.2 for COSINE SIMILARITY similarity and 0.15 for ARMORM, which were determined from the training set. Additionally, we evaluated the overlap coefficient (Vijaymeena and Kavitha, 2016) for detected privacy segment overlap using the following formula:$\text{OC}(A, B) = \frac{|A \cap B|}{\min(|A|, |B|)}$ which measures the number of common tokens between two sets. The results show that fine-tuning LLAMA2-7B achieves the best ROUGE-1 89.38% and OC 88.88%. With sufficient training data, the fine-tuned model performs well in span detection. On the other hand, metric-based methods require an appropriate cutoff threshold, making them less accurate compared to a fine-tuned model.

RQ4: Can our method effectively prevent reconstruction attacks for text rewriting? A critical concern is whether an adversary, knowing the algorithm and accessing rewritten texts, can infer

Method	PRIVACY_NL	ROUGE-1	PPL
ARMORM	93.02%	73.68%	151.83
PRIVACY_NLI	94.88%	30.00%	132.32
combined	88.55%	36.36%	99.23

Table 6: Comparison of Tree search discriminator function

original private information. Using the optimal reconstruction framework from (Tong et al., 2025), we compute ASR bounds at the token level for differing tokens in u and y, with token alignment for length discrepancies. The 3.07% ASR in both context-free and contextual Bayesian attacks indicates robust protection, supporting the adaptability of NaPaRe to diverse privacy specifications without sacrificing coherence or relevance.

Cost latency analysis for performance. As a practical complement to our accuracy results, we provide an *estimated* compute–cost and latency analysis using the same setup, together with a simple performance-per-unit-cost summary; see App. A.6. This analysis is intended only as an order-of-magnitude estimate to contextualize deployment trade-offs.

5 Related Work

Privacy in Large Language Models Recent research highlights growing concerns over privacy risks in large-scale language models, where both explicit and implicit private information can be inferred from text generation (Brown, 2020; Dou et al., 2024). Attack methods such as membership inference (Shokri et al., 2017) and reconstruction attacks (Lukas et al., 2023) reveal that models can memorize and leak sensitive details from training data. Prior studies have explored differential privacy mechanisms (Igamberdiev and Habernal, 2023; Bo et al., 2019), adversarial training (Barrett et al., 2019), and explicit text anonymization (Akbik et al., 2019; Lison et al., 2021) to mitigate these risks given various context, but these methods often degrade text utility or limited in the certain form of privacy requirements.

Privacy-Preserving Text Rewriting Privacy-aware text rewriting approaches typically rely on rule-based scrubbing (Akbik et al., 2019), fine-tuned anonymization models (Dou et al., 2024), or zero-shot prompting techniques (Utpala et al., 2023). Rule-based approaches are precise but limited in flexibility, while fine-tuned models require extensive human supervision. Recent work has leveraged prompt engineering for privacy preservation without retraining models, demonstrating effectiveness in document rewriting (Meisenbacher et al., 2024). Staab et al. (2024) considers LLM as an adversary to give feedback to the rewrite model to minimize the re-identification risk. Our study goes another direction to rewrite via tree-based rewrite with an explicit rewrite strategy.

Tree Search and Reward-Guided Generation Tree search techniques have been increasingly explored in LLM-controlled text generation, allowing structured decision-making over multiple reasoning and generation paths. Tree of Thoughts (ToT) enables LLMs to explore multiple reasoning paths systematically, improving complex task-solving by iterating over various candidate solutions (Yao et al., 2023). Similarly, Self-Play with Tree Search Refinement (SPaR) enhances model instruction by refining generated outputs through structured search and iterative decision-making (Cheng et al., 2024). Our method builds upon these principles by tree search and iterative refinement for privacy-preserving rewriting, ensuring progressive modifications that balance privacy, naturalness, and semantic preservation.

6 Conclusion

This paper introduces NaPaRe, a zero-shot tree-search based iterative privacy-aware rewriting method that adopts a MCTS-inspired search strategy. Through our extensive experiments, we show that our approach significantly outperforms baselines in terms of improved preservation of privacy, utility and naturalness. The MCTS-inspired search strategy is also superior to alternative methods. One possible direction is to further adapt and optimize NaPaRe as a versatile and generalized user privacy rewrite solution, particularly for on-device LLMs, to better accommodate evolving data release scenarios and granular user preferences.

7 Limitations

While NaPaRe effectively removes sensitive information and improves controllability, it has several limitations. First, the approach relies on model-driven rewriting, which may still retain implicit privacy cues or introduce inconsistencies due to the inherent variability of zero-shot prompting. Additionally, our method primarily focuses on general

textual data, but privacy risks vary across formats such as emails, chat messages, and structured documents. Expanding the framework to context-aware privacy preservation could improve adaptability across different communication settings.

Second, due to budgetary constraints, our method's implementation was limited in scope, preventing large-scale human annotation for diverse rewriting strategies. While the dataset is sufficient to validate our findings, it may not generalize to all real-world privacy scenarios, particularly in commercial settings. Future work could explore generalized expansion via prompt tuning or more efficient algorithms to reduce computational costs. Moreover, our evaluation primarily relies on automatic metrics. Developing more refined privacy evaluation metrics that better align with human preferences presents a promising direction for future research.

References

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*.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. 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 6331–6336.

Haohan Bo, Steven HH Ding, Benjamin Fung, and Farkhund Iqbal. 2019. Er-ae: differentially-private text generation for authorship anonymization. *arXiv preprint arXiv:1907.08736*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Jiale Cheng, Xiao Liu, Cunxiang Wang, Xiaotao Gu, Yida Lu, Dan Zhang, Yuxiao Dong, Jie Tang, Hongning Wang, and Minlie Huang. 2024. Spar: Self-play with tree-search refinement to improve instruction-following in large language models. *Preprint*, arXiv:2412.11605.

Nicola Dainese, Matteo Merler, Minttu Alakuijala, and Pekka Marttinen. 2024. Generating code world models with large language models guided by monte carlo tree search. *arXiv preprint arXiv:2405.15383*.

Dominic Deuber, Michael Keuchen, and Nicolas Christin. 2023. Assessing anonymity techniques employed in german court decisions: A {De-Anonymization} experiment. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5199–5216.

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, Bangkok, Thailand. Association for Computational Linguistics.

Minxin Du, Xiang Yue, Sherman SM Chow, Tianhao Wang, Chenyu Huang, and Huan Sun. 2023. Dp-forward: Fine-tuning and inference on language models with differential privacy in forward pass. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 2665–2679.

Shuo Huang, William MacLean, Xiaoxi Kang, Anqi Wu, Lizhen Qu, Qiongkai Xu, Zhuang Li, Xingliang Yuan, and Gholamreza Haffari. 2024. Napˆ 2: A benchmark for naturalness and privacy-preserving text rewriting by learning from human. *arXiv preprint arXiv:2406.03749*.

Timour Igamberdiev and Ivan Habernal. 2023. Dp-bart for privatized text rewriting under local differential privacy. *arXiv preprint arXiv:2302.07636*.

Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer.

Pierre Lison, Ildikó Pilán, David Sánchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. 2024. Dp-mlm: Differentially private text rewriting using masked language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9314–9328.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2023. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

NVIDIA Corporation. 2022. NVIDIA RTX-6000 GPU. https://www.nvidia.com/en-au/design-visualization/rtx-6000/.

NVIDIA Corporation. 2026. NVIDIA GeForce RTX 4070 Ti. https://www.nvidia.com/en-au/geforce/graphics-cards/50-series/rtx-5070-family/.

Lei Pan, Yunshi Lan, Yang Li, and Weining Qian. 2024. Unsupervised text style transfer via llms and attention masking with multi-way interactions. *arXiv preprint arXiv:2402.13647*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

shunzh. 2024. Monte-carlo tree search for large language models (mcts-for-llm). https://github.com/shunzh/mcts-for-llm. Commit 0785eda83e452be318780003c5c1b9821debfbdc, accessed 2025-09-08.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

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

Meng Tong, Kejiang Chen, Xiaojian Yuan, Jiayang Liu, Weiming Zhang, Nenghai Yu, and Jie Zhang. 2025. On the vulnerability of text sanitization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5150–5164.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. 2023. Locally differentially private document generation using zero shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8442–8457.

MK Vijaymeena and K Kavitha. 2016. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3(2):19–28.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *Preprint*, arXiv:2406.12845.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024b. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

A appendix

A.1 Baseline Method Implementation

DP-PROMPT. Utpala et al. (2023) utilizes zero-shot prompting and large language model to generate document paraphrasing to prevent author deanonymization attack which comprise the privacy of text owner. In our tasks, as our backbone model is LLAMA3.1-8B, we also use it as DP-PROMPT backbone. For ϵ, we set it to 100 which is best empirical to balance all metric tested.

DP-MLM Meisenbacher et al. (2024) considers BERT and Masked Language Prediction to gather improve the word utility for the generation distribution which bring improved utility in resulting generation. We maintain the same ϵ level to 100 to have comparative with other method.

DISCLOSURE-ABSTRACTION Dou et al. (2024) The self-disclosure abstraction detects the self disclosure span within given input text and perform rewrite using finetuned LLAMA3.1-8B model. As their defined privacy is close to the one defined in NAP^2, we directly adapts the rewrite model and test the generated result via our metric.

GPT4 The most powerful commercial language model(Achiam et al., 2023), we used the same prompt template from (Huang et al., 2024) to generate strategy specific rewrite as one of the grouding baseline for our method.

T5-NAP^2. The SOTA privacy rewriting model based on T5-BASE. It is fine-tuned on the dataset of NAP^2 directly with original utterance and human rewrite yielding the best privacy preserving score among all method

A.2 Upper Confidence Bound for Trees

To guide the exploration of candidate rewrites during our NaPaRe, we adopt the Upper Confidence Bound for Trees (UCT) algorithm. UCT balances exploitation of high-reward candidates with exploration of less-visited options by selecting actions that maximize the following objective:

$$\text{UCT}(i) = \bar{X}_i + C \cdot \sqrt{\frac{\ln N}{n_i}},$$

where $\bar{X}_i$ is the average reward of node i, n_i is the number of times node i has been visited, N is the total number of visits to the parent node, and C is a tunable exploration constant. In our setting, this mechanism allows us to prioritize rewrite trajectories that yield high reward scores while still exploring diverse rewriting paths. We set C empirically as 6.36 based on validation performance to ensure sufficient exploration during tree expansion.

A.3 Implementation Detail

We conducted our experiments on a single A40 GPU(NVIDIA Corporation, 2022) with 46GB RAM, ensuring efficient execution of our tree-search-based privacy rewriting method. And we also tested that our method can be run in GTX5070ti with 11GB RAM(NVIDIA Corporation, 2026).

The implementation of NaPaRe is adapted from the open-source repository Our MCTS decoder is adapted from the open-source *mcts-for-llm*[1] implementation (shunzh, 2024). , released under MIT. We adopted the frame for MCTS decoding and convert it to sentence level generation and implement our algorithm based on the framework. To control the computational overhead, we set the tree search computation budget to 5, allowing iterative refinement while maintaining feasible inference times. For model generation, we adopted a top-p probability of 5, ensuring diverse sampling while maintaining high-quality outputs. The maximum generation length was constrained to 128 tokens to prevent excessive expansion and maintain sentence coherence.

We set threshold for our reward model empirically to 0.10 to filter the rewrite quality. This threshold is obtained via training set to obtain best performance. Inference on our test set (280 instances) required approximately 2 hours, while processing the ECHR dataset took significantly longer, requiring 23 hours due to the complexity of legal text and entity alignment. In our experiment of ECHR dataset, the consumption is significant for some reasons. As the ECHR dataset is constructed by cases which contains 109 sentences per case with 20.33 words per sentence. By contract, each example in NAP is just one sentence with 140 examples with 14.25 tokens per examples which is more natural for human conversations. Thus the long running time for ECHR is acceptable under this circumstances in our experiment. It is noted that the generation with more inference steps raises significant computation overhead and latency. It

[1]The code is available at https://github.com/shunzh/mcts-for-llm.

also raises the further direction of optimizing such generation approach to minimize the inference cost.

A.4 Limitation for LLMs

In our work, the spotted weakness for privacy rewrite of LLM in one-off rewrite particular happens in the scenario where the sentence requires rewrite based on persona, for example if user ask to rewrite sentence "I am an Ohio Mon with two amazing sons, not married though" based on persona "I am a single mom of two boys". The SOTA LLM often fails to rewrite all possible private information mentioned in the persona as their alignment have different focus in multiple target. For GPT-4 model, the rewrite goes to "I live in Ohio and have a wonderful family, but I'm not married" and for open source model which is less competitive the result goes worse while handling complex privacy rewrite. Thus, we argue to decompose the private information for sentence to obtain better rewrite.

A.5 Additional Utility Metrics

Scope. These measurements provide supplementary evidence about text variety and semantic alignment. They are *complements* to our primary utility/accuracy metrics and should be interpreted accordingly.

Definitions. (1) *Diversity (Distinct-2)* = #unique bigrams / #total bigrams (computed per sample and averaged); (2) *MAUVE* measures distributional similarity between the model rewrites and references (higher is better); (3) *SimCSE* is cosine similarity between sentence embeddings of rewrite and reference (higher is better).

Setup. All methods are evaluated on the same test set and decoding configuration as the main experiments; preprocessing and tokenization follow the respective reference implementations of each metric.

Notes and caveats. Diversity and MAUVE can be sensitive to generation length/temperature; SimCSE depends on the encoder backbone. These indicators are provided to triangulate quality, not to replace task-specific utility or privacy outcomes.

A.6 Estimated Cost–Latency–Performance Analysis

Goal. We estimate how our method trades off privacy performance against compute cost and latency under the same evaluation setup. This is *not* a billing statement; all numbers are approximate and depend on hardware, rates, batching, and provider pricing.

Definitions. Let $P_{\text{ours}}, P_{\text{base}}$ be privacy scores on the test set (higher is better), and $C_{\text{ours}}, C_{\text{base}}$ be the estimated compute cost per 100 examples under identical conditions. We report the *incremental performance gain* $\Delta P = P_{\text{ours}} - P_{\text{base}}$ and the *performance-per-unit-cost*

$$\frac{\Delta P}{\Delta C} = \frac{P_{\text{ours}} - P_{\text{base}}}{C_{\text{base}} - C_{\text{ours}}}$$

This normalizes incremental benefit by incremental (net) cost on the same task, which is appropriate when contrasting alternative implementations of the *same* functionality.

Concrete figures. We use LLAMA 3.1–8B on an A40 GPU. Processing 100 examples takes **42.5 minutes** (≈ **51 s**/sentence). At a *nominal* rental rate of \$**0.47**/hour, this yields $C_{\text{ours}} = \$\,\mathbf{0.332}$ (=0.708 h $\times$ \$$0.47$/h). For a one-pass GPT-4 baseline via API we use $C_{\text{base}} = \$\,\mathbf{0.42}$ per 100 examples. Measured privacy scores are $P_{\text{ours}} = \mathbf{93.02}$ and $P_{\text{base}} = \mathbf{82.24}$, so $\Delta P = \mathbf{10.78}$. Therefore,

$$\frac{\Delta P}{\Delta C} = \frac{10.78}{0.42 - 0.332} = \frac{10.78}{0.088} = \mathbf{122.5}$$

Scope and caveats. All values are estimates; real costs vary with GPU market rates, region, utilization, batch size, tokenization, prompt length, caching, parallelization, and API pricing tiers. We exclude engineering time, storage, networking, and overheads. Latency reflects a single-GPU desktop/offline configuration; cloud inference, multi-GPU parallelism, or quantization can materially change results. We encourage readers to recompute with their own rates using the formulas above.

Reproduction details. For open-weight inference we compute $C_{\text{ours}} = r \times t$ with $r = \$0.47$/h and $t = 0.708$ h for 100 examples; for API we use a per-100-example estimate of \$0.42 aligned to our prompt/response lengths. Please substitute your own r and token pricing to recompute locally.

A.7 Detailed explanation of the example

We consider the rewrite example with input sentence "I am an Ohio Mon with two amazing sons, not married though" and persona information "I am a single mom of two boys." As shown in Figure. 2. From the datasets. There will be four parts of token considered as private segments based on

Method	Diversity (Distinct-2)	MAUVE	SimCSE
DP-MLM	2.5882	0.2391	0.4888
DP-PROMPT	2.9810	0.0044	0.1719
FLAIR-SCRUBBING	1.1999	0.0170	0.6572
DISCLOSURE-ABSTRACTION	2.0976	0.0725	0.2987
GPT4 (one-pass)	1.3521	**0.5878**	**0.5691**
T5-NAP^2-GPT4	0.5932	0.1794	0.2696
NaPaRe-LLAMA3.1-8B	**2.1293**	0.0235	0.5225

Table 7: Additional utility metrics (higher is better). Diversity is Distinct-2; SimCSE is cosine similarity to the reference. MAUVE compares the distributions of references and rewrites.

Method	P	ΔP	Time (100 ex.)	Cost/100 ex.	ΔC	$\Delta P/\Delta C$
Ours (LLAMA 3.1–8B, A40)	93.02	10.78	42.5 min	\$0.332	0.088	122.5
Baseline (GPT-4, one-pass)	82.24		–	\$0.42		

Table 8: Estimated cost–latency–performance summary for 100 examples. $\Delta P/\Delta C$ is in *points per dollar*. Figures are approximate and for contextual comparison only.

persona. In this case we conduct token-level mapping ['mom' 'two' 'sons' 'Not married'] For each private segments, we conduct our sentence level privacy rewrite in tree search with either obscuring or deleting strategies. The node randomly takes action of deleting at the beginning resulting in the output "I am an Ohio with two amazing sons, not married though". The reward function evaluates the rewrite and decides to take the next action of deleting in original input sentence. The subsequent output "I am an Ohio resident with two amazing sons, not married" are chosen via empirically set threshold for 0.10. It is accepted as partial rewritten sentence for next privacy segment. After that the next private segment "two" will be deleted. After all rewrites, the final output goes to "I am an Ohio resident with amazing children".

A.8 ARMORM

We adopt ARMORM (Wang et al., 2024b) as the reward model guiding our tree-search-based rewriting process. ARMORM is a state-of-the-art preference modeling framework that learns multi-objective reward functions from human feedback through a mixture-of-experts design. It produces interpretable reward signals aligned with human judgment across multiple dimensions, such as coherence, relevance, and instruction-following quality. We select ARMORM due to its strong empirical performance in evaluating text generation quality and its modular design that allows fine-grained control over different objectives during rewriting. They incorporate the Mixture-of-Experts (MoE) gating mechanism that selects objective-specific weights dynamically depending on the context (prompt), allowing adaptive and steerable reward aggregation. In our setting, the reward score generated by ARMORM is used to evaluate each candidate rewrite during the tree search process. Emperical result in comparsion shows it works

A.9 LLM as a judge evaluation of naturalness

We employ GPT-4o as LLM judge to score the generated output. As shown in Table. 1, each example using template in Fig. 3. The score ranges from 1 to 5. 1 denotes very unnatural of result and 5 denotes the high fluent and native sounding of a rewritten sentences. We set the human rewrite as naturalness to set comparison baseline for all methods.

A.10 More examples with rewrite

Table 9 displays 8 privacy rewriting examples from two domains. Examples 1–4 are from the NAP^2, where each input sentence is accompanied by a user-provided persona sentence that conveys private specification. The rewriting task involves aligning the input and persona to identify segments for rewriting. These privacy spans are highlighted in bold within the sentence. The rewrites aim to generalize or obfuscate sensitive information while maintaining fluency and semantic coherence. In Example 3, we additionally show a human-authored rewrite for comparison. As per request that we assume user require to fully remove information from the sentence. In human judgment it requires fully remove of later part. In this case even information of disability is hided by model. It actually failed to

LLM Naturalness Judgment Prompt

You are an expert linguist. Your task is to assess the naturalness of a given sentence — how fluent, human-like, and typical it sounds in everyday language use.

Rate the sentence on a scale from 1 to 5:

- **1** = very unnatural (awkward, grammatically incorrect, or robotic)
- **2** = mostly unnatural
- **3** = somewhat natural (acceptable but slightly awkward)
- **4** = mostly natural (minor issues)
- **5** = very natural (fluent and native-sounding)

Sentence: `"Sentence to Assess"`

Only provide the score and a brief explanation in the following JSON format:

```
{"score": X, "explanation": "..."}
```

Figure 3: Prompt used for LLM-based naturalness judgment.

give the desired rewrite like human rewrite.

Examples 5–8 are randomly drawn from the ECHR dataset, comprising legal case descriptions with sensitive attributes such as nationality, date of birth, location, and organization names. These were automatically scrubbed using the FLAIR-SCRUBBING, with the masked spans treated as private segments for rewriting. The rewritten outputs illustrate the model's ability to reconstruct fluent, semantically accurate paraphrases while replacing private entities with abstract or generalized descriptions. For instance, in Example 6, organizational and location-specific terms are replaced with more generic descriptors of institutions and international settings.

NAP^2 *Example 1*
Sentence: **Scotch** is my drink . Helps keep the evil dogs away , they scare me .
Persona: I like to drink **scotch** to relax.
Rewrite: My amber liquid of choice. Helps keep the menacing beasts at bay, they frighten me.

NAP^2 *Example 2*
Sentence: Hello . I live in an **apartment** . It is a **low income** residence .
Persona: I live in **low income apartments**.
Rewrite: Hello. I reside in our community.

NAP^2 *Example 3*
Sentence: That is great . I am **unable** to work due to my **disability**.
Persona: I am unable to **work** at a **job** because of a **disability**.
Rewrite: "That is great. I am currently experiencing certain challenges in my ability to work due to my personal circumstances."
Human Rewrite: That is great!

NAP^2 *Example 4*
Sentence: I **did not graduate** from **high school**, so writing novels seems like a great achievement
Persona: I **never finished high school**.
Rewrite: I achieved great things despite completing my academic journey, so crafting stories feels like a remarkable accomplishment.

ECHR *Example 5*
Sentence: The applicant is a **British national**, born in **1945** and living in **Rome**.
Masked: The applicant is a <MASK> national, born in <MASK> and living in <MASK>.
Rewrite: The applicant is a native of a major European capital city, born during the post-war era and residing in the Eternal City.

ECHR *Example 6*
Sentence: In **1982** the **first** applicant, a civil engineer by profession and employed by the **French** company S., was placed at the disposal of **the European Space Agency** to perform services at the **European Space Operations Centre** in **Darmstadt**.
Masked: In <MASK> the <MASK> applicant, a civil engineer by profession and employed by the <MASK> company S., was placed at the disposal of <MASK> to perform services at <MASK> in <MASK>.
Rewrite: "A trailblazing innovation specialist, with a background in urban development, was seconded to a prestigious intercontinental space research institution to provide expertise in a cosmopolitan hub.

ECHR *Example 7*
Sentence: After repeated motions for bias filed by the applicant's husband, the **Buda Central District Court** was **first** appointed to hear the case.
Masked: After repeated motions for bias filed by the applicant's husband, <MASK> was <MASK> appointed to hear the case.
Rewrite: "After repeated motions for bias filed by the applicant's husband, a court was appointed to hear the case."

ECHR *Example 8*
Sentence: By **two** separate judgments of **4 June 2007**, **one** in favour of Mr **Pavlenko** and another in favour of the remaining applicants, the **Mineralniye Vody Town Court** of the **Stavropol Region** upheld their actions in part.
Masked: By <MASK> separate judgments of <MASK>, <MASK> in favour of Mr <MASK> and another in favour of the remaining applicants, <MASK> of <MASK> upheld their actions in part.
Rewrite: Copies were sent to nearly two dozen other individuals in the government and the federal authorities and to the diplomatic missions in the coastal metropolis, the radiant capital, and the central city.

Table 9: 8 randomly sampled examples generated with NaPaRe-LLAMA3.1-8B. 4 examples from NAP^2 and 4 examples from ECHR. The privacy segments are marked **bold**