Mask-Free Privacy Extraction and Rewriting: A Domain-Aware Approach via Prototype Learning

Xiaodong Li[1,2*], **Yuhua Wang**[2,3*], **Qingchen Yu**[2,3], **Zixuan Qin**[1], **Yifan Sun**[1,2†], **Qinnan Zhang**[2,3†], **Hainan Zhang**[2,3], **Zhiming Zheng**[2,3]

[1]Center for the Applied Statistics, School of Statistics, Renmin University of China
[2]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing
[3]School of Artificial Intelligence, Beihang University, China

muzi_4@ruc.edu.cn, yuhuawang@buaa.edu.cn

Abstract

Client-side privacy rewriting is crucial for deploying LLMs in privacy-sensitive domains. However, existing approaches struggle to balance privacy and utility. Full-text methods often distort context, while span-level approaches rely on impractical manual masks or brittle static dictionaries. Attempts to automate localization via prompt-based LLMs prove unreliable, as they suffer from unstable instruction following that leads to privacy leakage and excessive context scrubbing. To address these limitations, we propose **DAMPER** (**D**omain-**A**ware **M**ask-free **P**rivacy **E**xtraction and **R**ewriting). DAMPER operationalizes latent privacy semantics into compact **Domain Privacy Prototypes** via contrastive learning, enabling precise, autonomous span localization. Furthermore, we introduce a Prototype-Guided Preference Alignment, which leverages learned prototypes as semantic anchors to construct preference pairs, optimizing a domain-compliant rewriting policy without human annotations. At inference time, DAMPER integrates a sampling-based Exponential Mechanism to provide rigorous span-level Differential Privacy (DP) guarantees. Extensive experiments demonstrate that DAMPER significantly outperforms existing baselines, achieving a superior privacy-utility trade-off.[1]

1 Introduction

Deploying cloud-hosted Large Language Models (LLMs) in privacy-sensitive domains (e.g., healthcare, law, finance) often necessitates transmitting sensitive queries (Qin et al., 2025; Yu et al., 2025), posing privacy risks under untrusted providers (Staab et al., 2024; Kim et al., 2023).

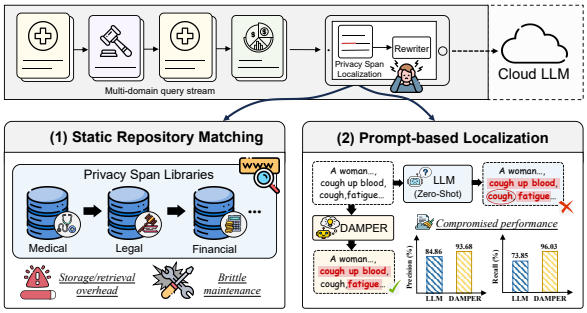


Figure 1: **Challenge illustration.** (1) Static repository matching requires distinct privacy libraries for each domain, incurring excessive storage and maintenance overheads. (2) Prompt-based localizations lack granularity, resulting in compromised performance: they indiscriminately mask generic terms (low Precision) while missing actual sensitive ones (suboptimal Recall). Conversely, DAMPER leverages domain prototypes to discern context, achieving superior performance.

Client-side privacy rewriting mitigates this by locally sanitizing content while preserving task intent (Shi et al., 2022; Igamberdiev et al., 2022).

Early approaches (Mattern et al., 2022; Utpala et al., 2023; Meisenbacher et al., 2024) primarily operated at the *full-text level*, rewriting the entire input and uniformly applying Differential Privacy (DP) mechanisms (Dwork, 2006). While straightforward, such coarse-grained perturbations often distort non-sensitive context, eroding domain style and professionalism and degrading downstream utility. This has motivated *span-level privacy rewriting* (Huang et al., 2025; Zeng et al., 2025), which selectively rewrites only privacy spans—contiguous segments that encode sensitive attributes—while preserving the remaining context verbatim. However, many span-level methods **implicitly assume that users can provide privacy masks or otherwise specify which spans are sensitive**. This assumption rarely holds in practical client deployment, where query streams span multiple professional domains and privacy semantics are

*Equal contribution.
†Corresponding author.
[1]Code and a video demonstration are available at: https://github.com/muzi-4/DAMPER-main.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 2987–3016
October 24-29, 2026 ©2026 Association for Computational Linguistics

inherently domain-contingent. Although users typically know what they want to ask, they often lack the expertise to localize privacy spans according to domain-specific standards. Therefore, a practical rewriter must autonomously infer domain context and localize privacy spans, a capability we term *mask-free* rewriting.

Two natural strategies for mask-free localization fall short in practice. ❶ **Static Repository Matching.** Supporting multi-domain streams requires maintaining separate privacy span libraries, incurring significant storage and retrieval overheads. Moreover, the open-ended nature of sensitive vocabulary makes such maintenance brittle and costly. ❷ **Prompt-based Localization.** Even when provided with detailed prompts strictly aligned with the privacy definitions used for dataset annotation (see Appendix I.6), zero-shot local LLMs struggle with span-level instruction following. As detailed in our case studies in Appendix H, this instability leads to simultaneous false positives and negatives. Such errors either *over-scrub* benign context, thereby degrading downstream utility, or leave privacy spans *exposed*, directly compromising user privacy. Ultimately, this corrupts the evidentiary basis required for reliable downstream reasoning.

These limitations call for a lightweight, domain-aware framework capable of capturing latent privacy concepts to enforce domain-consistent boundaries. This setting, where both domain semantics and privacy boundaries are latent at inference time, raises two fundamental questions: **Q1.** *How can we operationalize domain-specific privacy semantics to enable autonomous span localization without manual masks?* **Q2.** *How can we learn a utility-preserving rewriting policy without human annotations, while strictly satisfying DP constraints?*

In this work, we propose **D**omain-**A**ware **M**ask-Free **P**rivacy **E**xtraction and **R**ewriting, abbreviated as **DAMPER**. DAMPER operationalizes domain privacy semantics with **Domain Privacy Prototypes**—compact latent representations that guide the full lifecycle from localization to rewriting. Our key intuition is that ***domain matters twice***: it determines ***which*** spans are privacy-sensitive, and it constrains ***how*** they should be rewritten to maintain professional consistency. To address **Q1**, we introduce a prototype-driven localization mechanism. We learn discriminative span representations using a multi-domain contrastive objective that pulls together private spans within the same domain while pushing apart non-private text and private spans from other domains. We then cluster these embeddings to obtain Domain Privacy Prototypes, which compress diverse and sensitive instances into a small set of semantic anchors. At inference time, DAMPER computes affinities between input segments and the learned prototypes to infer the dominant domain and localize privacy spans *without* manual masks. To tackle **Q2**, we propose a preference-based rewriting paradigm that explicitly manages the tension between obfuscation and context preservation. Rather than imitating static references, we employ Prototype-Guided Direct Preference Optimization (DPO) (Ouyang et al., 2022) with automatically constructed preference pairs that penalize rewrites overly similar to the original sensitive span (privacy) while rewarding consistency with the domain prototype (utility). This yields a domain-compliant rewriting policy *without* human preference annotations. During inference, we integrate a sampling-based Exponential Mechanism into token generation to provide span-level DP guarantees. Our contributions are:

- We propose a domain-aware rewriting framework that decouples sanitization from downstream LLMs, facilitating autonomous, mask-free span localization across multi-domains.
- We introduce a prototype-guided preference alignment paradigm that leverages learned prototypes as consistency anchors, explicitly navigating the conflict between obfuscation and preservation in an annotation-free manner.
- We derive a span-level DP guarantee via the Exponential Mechanism, and demonstrate through extensive experiments that DAMPER achieves superior privacy–utility trade-offs.

2 Related Work

Token-level LDP Substitution. Early approaches (Yue et al., 2021; Chen et al., 2023a) operate under metric local differential privacy, replacing each word with a neighbor in the embedding space. While providing formal guarantees, these methods treat every token as equally sensitive. This indiscriminate perturbation often destroys syntactic structure and semantic utility, as noise is injected regardless of the token's privacy contribution.

Sequence-level DP Paraphrasing. To improve coherence, sequence-level methods (Mattern et al., 2022; Igamberdiev and Habernal, 2023; Utpala et al., 2023; Meisenbacher et al., 2024) rewrite entire texts using encoder-decoder architectures or

LLM prompting. Although they generate more fluent outputs than token substitution, they apply a uniform privacy budget across the input. This lack of granularity forces a trade-off: sufficient noise to protect sensitive details often obfuscates the general context necessary for downstream tasks.

Span-level Privacy Rewriting. Recent frameworks adopt a detect-then-sanitize paradigm, utilizing strategies such as server-side restoration (Zeng et al., 2025), zero-shot rewriting (Huang et al., 2025), or inference-time distribution mixing (Thareja et al., 2025). However, these methods predominantly rely on rigid detection mechanisms (e.g., static inventories or explicit user masks), lacking the autonomous identification capability essential for open-ended domain semantics. Structurally, they often face a dilemma: search-based approaches typically lack formal DP guarantees, while restoration-dependent methods are tightly coupled to specific server-side LLMs, limiting model-agnostic deployment.

Full discussions of related work and preliminaries are deferred to Appendix B and C, respectively.

3 Problem Formulation

Threat Model. We consider a client-side setting where the user's device runs a local privacy rewriter before sending a query to an untrusted cloud LLM. The adversary is the cloud provider, or any party observing the transmitted query, with full access to the rewritten text y and arbitrary inference or auxiliary knowledge to recover sensitive content from the original input x. In particular, the adversary may try to reconstruct privacy spans or infer private attributes implied by them.

Span Definitions. We view a user query x as a sequence of tokens containing a set of semantically meaningful spans, denoted as $\mathcal{U}$. Within this universal set, we distinguish a subset of privacy spans $\mathcal{S} \subset \mathcal{U}$ that convey sensitive information (e.g., detailed clinical symptoms). The remaining segments in $\mathcal{U}$ are considered non-privacy spans—meaningful but explicitly non-sensitive phrases. Any tokens not belonging to $\mathcal{U}$ (e.g., functional connectives or generic terms) are treated as ordinary context. For example, in the medical text "*A woman . . . cough, cough up blood, fatigue . . .*", the set $\mathcal{U}$ includes meaningful segments like *cough*, *cough up blood*, and *fatigue*. Here, $\mathcal{S}$ consists of *cough up blood* and *fatigue*, while *cough* remains in $\mathcal{U}$ as a non-privacy span. Generic phrases like *A woman* fall outside $\mathcal{U}$ as ordinary context.

Task Definition. Given an input x, we aim to learn a rewriter π_θ that maps x to a rewritten query y. In open-world settings where both the target domain and privacy annotations $\mathcal{S}$ are latent, the proposed method must satisfy three objectives: 1) **Domain-Aware Localization**. Dynamically infer the domain context of x to localize the privacy spans $\hat{\mathcal{S}} \approx \mathcal{S}$ within $\mathcal{U}$, distinct from non-privacy spans. 2) **DP Rewriting**. Generate rewritten replacements for the detected spans $\hat{\mathcal{S}}$ under formal DP guarantees to obfuscate sensitive attributes. 3) **Utility Preservation**. Preserve the semantic integrity of the non-sensitive context ($\mathcal{U} \setminus \hat{\mathcal{S}}$) and the stylistic consistency of the domain, ensuring the rewritten query remains effective for downstream tasks.

4 Methodology

As shown in Figure 2, DAMPER operates in two phases. The Offline Training constructs Domain Privacy Prototypes to encode domain privacy and align the rewriting policy via preference learning. The Online Inference serves as a plug-and-play module for domain-adaptive span localization and differentially private rewriting of sensitive regions. Appendix E presents the algorithms of DAMPER.

4.1 Offline Training Phase

Privacy Prototype Generation. Privacy criteria exhibit inherent domain heterogeneity. Beyond generic PII, each domain adheres to distinct sensitivity standards, exemplified by disease severity in the medical field and crime classification in the legal domain. For each domain $k \in \mathcal{K}$, we collect annotated privacy spans from the training corpus $\mathcal{D}$ into $\mathcal{S}_k = \{s_{k,i}\}_{i=1}^{N_k}$. Ideally, these diverse spans could be summarized into compact prototypes via clustering. However, as shown in Figure 3, raw representations from pre-trained encoders are highly anisotropic and unstructured. Specifically, privacy and non-privacy spans across domains are entangled in the embedding space and lack discriminative boundaries for effective clustering.

To rectify this, we employ **multi-domain contrastive learning** to reshape the embedding manifold. Our objective is to enhance discriminability by encouraging intra-domain compactness among privacy spans while enforcing separability from non-privacy contexts and out-of-domain privacy spans. We instantiate a trainable span encoder $h(\cdot)$ by augmenting a frozen backbone $g(\cdot)$ with lightweight adapter layers. For an anchor privacy span $s_{k,i} \in \mathcal{S}_k$ with embedding $\mathbf{z}_{k,i} = h(s_{k,i})$, we

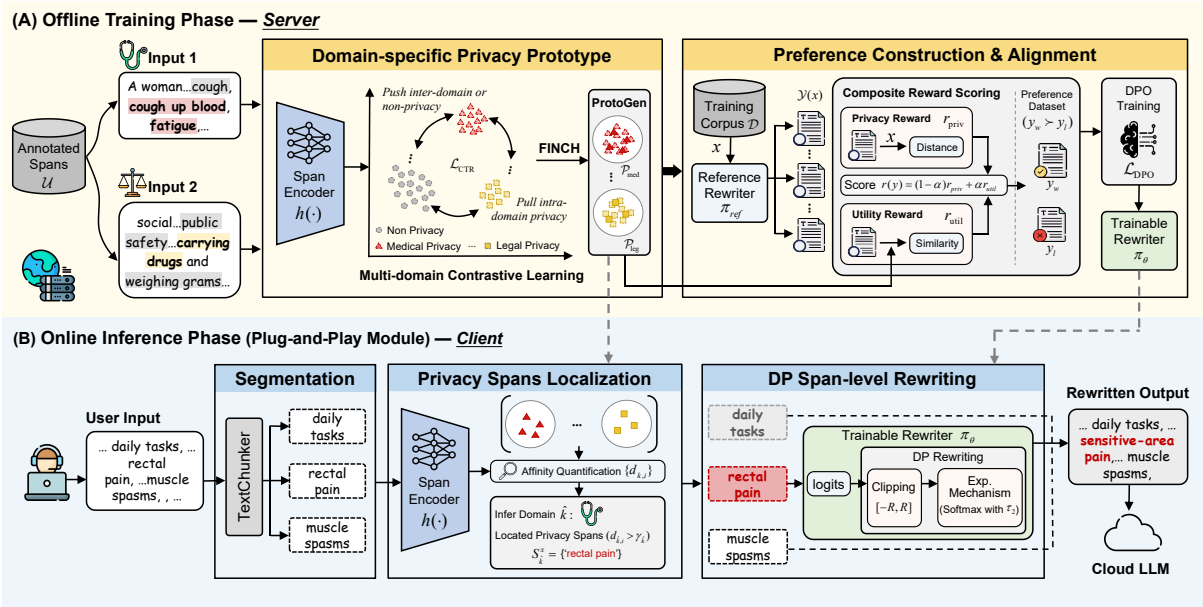
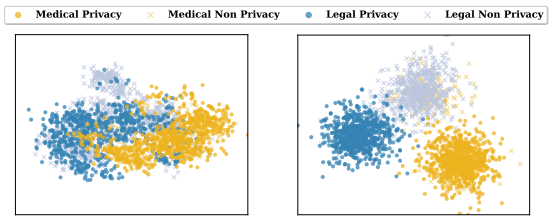


Figure 2: **Overview of the DAMPER**. (A) Offline Training Phase (Sec. 4.1): We first employ multi-domain contrastive learning to cluster annotated spans into compact *Domain Privacy Prototypes* (e.g., $\mathcal{P}_{med}$, $\mathcal{P}_{leg}$). These prototypes guide DPO training for rewriter π_θ using a composite reward to balance semantic obfuscation and domain fidelity. (B) Online Inference Phase (Sec. 4.2): The module segments user input via TextChunker and dynamically infers the target domain to localize privacy spans based on prototype affinity. Detected spans are rewritten using a sampling exponential mechanism, preserving non-sensitive context verbatim.



(a) Represented by $g(\cdot)$. (b) Represented by $h(\cdot)$.

Figure 3: **T-SNE visualization** of span representations produced by backbone $g(\cdot)$ vs. span encoder $h(\cdot)$.

construct the contrastive objectives based on:

- *Positive set* $\mathcal{A}_{k,i}$ consists of other privacy spans within the same domain k (i.e., $\mathcal{S}_k \setminus \{s_{k,i}\}$).
- *Negative set* $\mathcal{N}_{k,i}$ comprises privacy spans from disparate domains ($\mathcal{S}_j, j \neq k$) and all non-privacy spans (i.e., $\mathcal{U} \setminus \mathcal{S}$), which serve as hard negatives to enforce semantic discrimination.

We minimize the Multi-positive InfoNCE loss to learn discriminative span representations:

$$\mathcal{L}_{\text{CTR}} = -\log \frac{\sum_{\mathbf{a} \in \mathcal{A}_{k,i}} \exp\left(\cos(\mathbf{z}_{k,i}, \mathbf{a})/\tau_1\right)}{\sum_{\mathbf{g} \in \mathcal{G}_{k,i}} \exp\left(\cos(\mathbf{z}_{k,i}, \mathbf{g})/\tau_1\right)}, \quad (1)$$

where $\mathcal{G}_{k,i} = \mathcal{A}_{k,i} \cup \mathcal{N}_{k,i}$ encompasses all positive and negative samples, and τ_1 is the temperature.

Post-contrastive tuning, the encoder $h(\cdot)$ yields domain-aware representations. We then derive discrete privacy prototypes by applying the FINCH clustering algorithm (Sarfraz et al., 2019) to the refined privacy embeddings of each domain k:

$$\mathcal{P}_k = \{\mathbf{p}_{k,j}\}_{j=1}^{J_k} = \text{FINCH}\left(\{\mathbf{z}_{k,i}\}_{i=1}^{N_k}\right). \quad (2)$$

These prototypes serve as a structured abstraction of diverse in-domain privacy expressions, providing the foundation for our subsequent steps.

Preference Construction. To align the rewriter with both privacy protection and domain consistency, we propose a preference learning framework guided by the learned domain prototypes. Rather than training an explicit reward model, we construct a preference dataset with a composite reward function. For an input x with annotated privacy spans $\mathcal{S}_k^x$, we first generate a set of candidate rewrites $\mathcal{Y}(x)$ using a reference model π_{ref} (detailed in Appendix I.1). We then evaluate each candidate $y \in \mathcal{Y}(x)$ using two competing criteria. Let $\mathbf{z}_{k,i}^x = h(s_{k,i}^x)$ and $\mathbf{z}_{k,i}^y = h(s_{k,i}^y)$ denote the embeddings of the i-th original and rewritten privacy spans, respectively.

Semantic Obfuscation (Privacy). To ensure privacy, the rewritten spans should be semantically distant from the original sensitive information. We define the privacy reward $r_{\text{priv}}(y)$ as the negative cosine similarity between the embeddings of the rewritten

spans and the original spans:

$$r_{\text{priv}}(y) = 1 - \frac{1}{N_k^x} \sum_{i=1}^{N_k^x} \cos\left(\mathbf{z}_{k,i}^y, \mathbf{z}_{k,i}^x\right). \quad (3)$$

Domain Fidelity (Utility). To preserve domain style, rewritten spans must align with the latent structure of the target domain. We quantify utility $r_{\text{util}}(y)$ by measuring the proximity of each rewritten span to its nearest domain prototype:

$$r_{\text{util}}(y) = \frac{1}{N_k^x} \sum_{i=1}^{N_k^x} \max_j \cos\left(\mathbf{z}_{k,i}^y, \mathbf{p}_{k,j}\right). \quad (4)$$

Here, the $\max$ operator automatically selects the most relevant prototype for each span.

The final score is defined as

$$r(y) = (1-\alpha)\, r_{\text{priv}}(y) + \alpha\, r_{\text{util}}(y), \quad (5)$$

where $\alpha \in [0,1]$ is a trade-off hyperparameter. We leverage the composite reward to automatically construct a preference dataset $\mathcal{D}_{\text{pref}}$. For each input x, we identify the candidate with the highest score $r(y)$ as the preferred response y_w and the one with the lowest score as the dispreferred response y_l.

Preference Alignment. We fine-tune the rewriter π_θ using DPO, which optimizes the policy to align these synthesized preferences while constraining deviation from the reference model π_{ref}:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\Big(\beta \log \tfrac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \tfrac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big), \quad (6)$$

where β regulates the KL-divergence penalty. By focusing the optimization on privacy spans, this mechanism effectively achieves a granular balance between semantic privacy and domain style.

4.2 Online Inference Phase

We deploy the privacy rewriter as a plug-and-play module on the client side that adaptively rewrites queries prior to transmission to the cloud LLM.

Text Segmentation. Given a user query x, we first apply a rule-based segmenter, TextChunker$(\cdot)$, to decompose x into a sequence of semantically coherent spans $\{a_i\}_{i=1}^M$ based on punctuation and linguistic cues (detailed in Appendix D).

Privacy Span Localization. As the domain of user input is unknown at inference, we use the learned privacy prototypes to jointly infer the global domain and localize sensitive spans. Specifically, each span a_i is encoded as $\mathbf{z}_i = h(a_i)$. For a candidate domain k with prototype set $\mathcal{P}_k$, we define the affinity of span a_i to domain k by its maximum cosine similarity to the prototypes in $\mathcal{P}_k$:

$$d_{k,i} = \max_{1 \le j \le J_k} \cos(\mathbf{z}_i, \mathbf{p}_{k,j}). \quad (7)$$

We then aggregate these span-level affinities across the query and infer the global domain $\hat{k}$ as the one with the highest average affinity:

$$\hat{k} = \arg\max_{k \in \mathcal{K}} \frac{1}{M} \sum_{i=1}^{M} d_{k,i}. \quad (8)$$

Given the inferred domain $\hat{k}$, we identify a span a_i as private if its affinity $d_{\hat{k},i}$ exceeds a domain-specific threshold $\gamma_{\hat{k}}$. The resulting privacy span set is denoted by $\hat{\mathcal{S}}_{\hat{k}}^x$, while the remaining spans are treated as non-sensitive context.

To avoid manually setting $\gamma_{\hat{k}}$, we determine it adaptively using the maximum between-cluster variance criterion (Otsu, 1979). We first sort the affinities as $d_{\hat{k},(1)} \le d_{\hat{k},(2)} \le \cdots \le d_{\hat{k},(M)}$. For a split point t, let $\mu_l = \frac{1}{t}\sum_{i=1}^{t} d_{\hat{k},(i)}$, $\mu_r = \frac{1}{M-t}\sum_{i=t+1}^{M} d_{\hat{k},(i)}$, and $\mu = \frac{1}{M}\sum_{i=1}^{M} d_{\hat{k},(i)}$. The corresponding between-cluster variance is

$$\sigma^2(t) = \frac{t}{M}(\mu_l - \mu)^2 + \frac{M-t}{M}(\mu_r - \mu)^2. \quad (9)$$

We select the split point $t^\star = \arg\max_t \sigma^2(t)$ and set the threshold to the midpoint between the two adjacent sorted affinities:

$$\gamma_{\hat{k}} = \big(d_{\hat{k},(t^\star)} + d_{\hat{k},(t^\star+1)}\big)/2. \quad (10)$$

DP Span-level Rewriting. Given an input query x and detected privacy spans $\hat{\mathcal{S}}(x)$, our rewriter copies all non-private tokens verbatim and regenerates *only* tokens within $\hat{\mathcal{S}}(x)$. The privacy guarantee is therefore *span-restricted*: it protects changes within $\hat{\mathcal{S}}(x)$ while preserving the remaining context exactly. Prompt templates are in Appendix I.2.

At each decoding step ℓ for a private token, the rewriter outputs a logit vector $u_\ell \in \mathbb{R}^{|\mathcal{V}|}$ over the vocabulary $\mathcal{V}$. We bound the utility sensitivity by coordinate-wise clipping, $\bar{u}_\ell = \text{clip}(u_\ell; R_1, R_2)$, and sample the next token with a temperature-softmax distribution, which instantiates the Exponential Mechanism (McSherry and Talwar, 2007) with utility $\bar{u}_\ell(\cdot)$:

$$\Pr[w_\ell = w|x] = \frac{\exp\big(\bar{u}_\ell(w)/\tau_2\big)}{\sum_{w' \in \mathcal{V}} \exp\big(\bar{u}_\ell(w')/\tau_2\big)}. \quad (11)$$

Under clipping, each coordinate satisfies $\bar{u}_\ell(w) \in [R_1, R_2]$, hence the per-step utility sensitivity is

bounded by $\Delta u \triangleq \max_{x \sim x'} \max_{w \in \mathcal{V}} \left| \bar{u}_\ell(w; x) - \bar{u}_\ell(w; x') \right| \leq (R_2 - R_1)$, which yields a per-token privacy cost:

$$\epsilon_{\text{token}} = \frac{2\Delta u}{\tau_2} \leq \frac{2(R_2 - R_1)}{\tau_2}. \tag{12}$$

Let $n_{\text{sp}}(x)$ be the number of regenerated tokens within $\hat{\mathcal{S}}(x)$, and let $n_{\text{sp}}^{\max}$ be a global upper bound used for budgeting. By sequential composition over the *regenerated* tokens, the rewriting satisfies

$$\epsilon_{\text{text}}(x) \leq n_{\text{sp}}(x) \cdot \epsilon_{\text{token}} \leq n_{\text{sp}}^{\max} \cdot \epsilon_{\text{token}}. \tag{13}$$

We treat ϵ_{text} as the *total privacy budget for all regenerated tokens in a query*, and set the sampling temperature by inverting (12)–(13):

$$\tau_2 \triangleq \frac{2(R_2 - R_1)\, n_{\text{sp}}^{\max}}{\epsilon_{\text{text}}}. \tag{14}$$

This calibration ensures that the overall privacy loss scales with the amount of *sensitive* text being regenerated, rather than the full query length. Moreover, since $n_{\text{sp}}(x) \leq n_{\text{sp}}^{\max}$, the realized privacy loss is often strictly smaller than the budgeted ϵ_{text}.

Theorem 4.1 *Condition on the detected span set $\hat{\mathcal{S}}(x)$ and consider adjacent inputs $x \sim x'$ that differ only within token positions covered by $\hat{\mathcal{S}}(\cdot)$ (tokens outside $\hat{\mathcal{S}}(\cdot)$ are identical and copied deterministically). If private tokens are sampled by (11) with clipping bounds $[R_1, R_2]$ and temperature τ_2 set as in (14), then the randomized rewriting stage satisfies ϵ_{text}-DP with $\epsilon_{text} = n_{sp}^{\max} \epsilon_{token}$.*

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on two established datasets: Pri-DDXPlus (Zeng et al., 2025) (Medical) and Pri-SLJA (Zeng et al., 2025) (Legal). To assess domain perception and adaptability, we further construct a composite benchmark, Pri-Mixture, by aggregating samples from these base domains. Detailed statistics are provided in Appendix F.1.

Training Protocol. To demonstrate the framework's capability for automatic domain-aware privacy, *we train the model on Pri-Mixture to learn discriminative domain features* and evaluate it on Pri-DDXPlus, Pri-SLJA, and Pri-Mixture. This protocol challenges the model to dynamically adapt to varying contexts without explicit task indicators. Unless otherwise specified, all results report the performance of this unified model.

Baselines. We compare against No Rewriting (upper bound), DP-Paraphrase (Mattern et al., 2022), DP-Prompt (Utpala et al., 2023), DP-MLM (Meisenbacher et al., 2024), and PrivacyRestore (Zeng et al., 2025). To isolate localization effects, we introduce two analysis settings: (1) Oracle variants (DP-MLM$_{\text{oracle}}$, DAMPER$_{\text{oracle}}$), using ground-truth privacy spans to evaluate the rewriter in isolation. (2) DP-MLM$_{\text{auto}}$, a hybrid baseline that applies the DP-MLM rewriter specifically to spans detected by our prototype-guided localizer.

Metrics. We evaluate three dimensions: downstream utility via Accuracy (ACC), semantic consistency via BERTScore (BS), and generation quality via LLM-Judge (LLM-J) (Zheng et al., 2023; Zeng et al., 2025). Definitions and evaluation prompt templates are provided in Appendix F.2 and I.

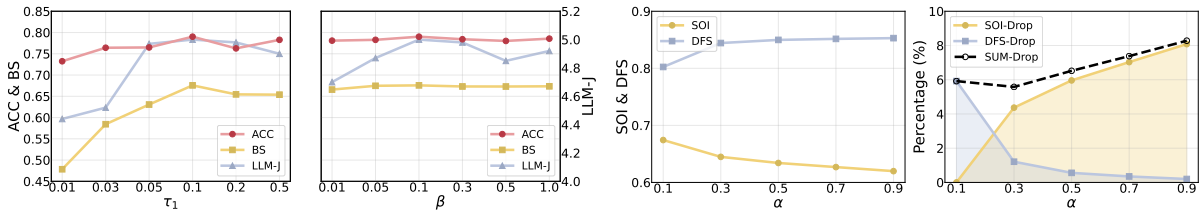
Implementation Details. We utilize RoBERTa-base (Liu et al., 2019) as the backbone for the span encoder and Qwen2.5-1.5B-Instruct (Team, 2024) as reference and policy models for preference learning. The cloud LLM is instantiated with Qwen2.5-7B-Instruct (Team, 2024; Yang et al., 2024). Default hyperparameters are configured as $\tau_1 = 0.1$, $\beta = 0.1$, and $\alpha = 0.3$. We adopt the same privacy hyperparameter ϵ from PrivacyRestore (Zeng et al., 2025). The ϵ_{token} of DAMPER is 2.88 on Pri-DDXPlus and 1.44 on Pri-SLJA. See Appendix F.3 for ϵ_{text}, ϵ_{token} and ϵ details.

5.2 Performance Comparison

Table 1 compares DAMPER against six baselines, showing superior or competitive performance across all metrics. ❶ In terms of downstream utility (ACC), DAMPER establishes a new state of the art. This advantage is pronounced on Pri-Mixture, where baselines suffer from inter-domain interference. DAMPER effectively preserves domain boundaries, confirming that our prototype-driven approach captures discriminative features robust to distractors. ❷ Regarding semantic consistency (BS), DAMPER leads on Pri-DDXPlus and Pri-Mixture, ranking second only to DP-MLM$_{\text{oracle}}$ on Pri-SLJA. This gap is expected, as the oracle uses ground-truth spans. Despite autonomous inference without masks, DAMPER maintains high semantic fidelity. ❸ For generation quality (LLM-J), DAMPER generally outperforms baselines. While PrivacyRestore shows a slight edge on Pri-SLJA under high ϵ, it requires expensive retrieval. Conversely, DAMPER achieves comparable utility through efficient one-pass rewriting without post-

Methods	Pri-DDXPlus			Pri-SLJA			Pri-Mixture		
	ACC ↑	BS ↑	LLM-J ↑	ACC ↑	BS ↑	LLM-J ↑	ACC ↑	BS ↑	LLM-J ↑
No Rewriting	$87.60_{\pm 0.07}$	$1.00_{\pm 0.00}$	$5.60_{\pm 0.14}$	$89.79_{\pm 0.28}$	$1.00_{\pm 0.00}$	$6.89_{\pm 0.08}$	$88.45_{\pm 0.14}$	$1.00_{\pm 0.00}$	$6.04_{\pm 0.12}$
DP-Paraphrase	$45.35_{\pm 0.71}$	$0.44_{\pm 0.00}$	$1.66_{\pm 0.02}$	$46.86_{\pm 0.97}$	$0.41_{\pm 0.00}$	$1.94_{\pm 0.04}$	$43.06_{\pm 0.33}$	$0.43_{\pm 0.00}$	$1.76_{\pm 0.03}$
DP-Prompt	$29.56_{\pm 0.57}$	$0.28_{\pm 0.00}$	$1.68_{\pm 0.01}$	$23.97_{\pm 0.83}$	$0.23_{\pm 0.00}$	$2.14_{\pm 0.16}$	$27.09_{\pm 0.90}$	$0.27_{\pm 0.00}$	$1.84_{\pm 0.06}$
DP-MLM	$33.12_{\pm 0.39}$	$0.39_{\pm 0.00}$	$1.70_{\pm 0.06}$	$25.93_{\pm 2.91}$	$0.28_{\pm 0.00}$	$2.22_{\pm 0.04}$	$28.93_{\pm 1.25}$	$0.35_{\pm 0.00}$	$1.88_{\pm 0.03}$
DP-MLM$_{\text{oracle}}$	$50.94_{\pm 0.88}$	$\underline{0.59}_{\pm 0.00}$	$3.61_{\pm 0.10}$	$\underline{78.78}_{\pm 0.69}$	$\mathbf{0.73}_{\pm 0.00}$	$4.81_{\pm 0.02}$	$59.60_{\pm 0.91}$	$\underline{0.64}_{\pm 0.00}$	$4.02_{\pm 0.06}$
PrivacyRestore	$\underline{75.82}_{\pm 0.17}$	-	$\underline{4.32}_{\pm 0.01}$	$78.38_{\pm 1.02}$	-	$\underline{5.26}_{\pm 0.03}$	$\underline{76.35}_{\pm 0.56}$	-	$\underline{4.64}_{\pm 0.02}$
DAMPER	$\mathbf{78.13}_{\pm 1.13}$	$\mathbf{0.68}_{\pm 0.00}$	$\mathbf{4.76}_{\pm 0.21}$	$\mathbf{82.68}_{\pm 0.86}$	$\underline{0.66}_{\pm 0.00}$	$\mathbf{5.37}_{\pm 0.04}$	$\mathbf{78.29}_{\pm 0.59}$	$\mathbf{0.67}_{\pm 0.00}$	$\mathbf{4.97}_{\pm 0.09}$

Table 1: Performance comparison on three datasets. ACC is reported in %, where "-" denotes unavailable results. Results are averaged over 3 runs. The **best** and second are marked.



(a) The effect of τ_1. (b) The effect of β. (c) The effect of α. (d) The effect of α.

Figure 4: **Hyper-parameter sensitivity** of DAMPER on Pri-Mixture.

$\mathcal{L}_{CTR}$	$\mathcal{L}_{DPO}$	Pri-Mixture ACC ↑	BS ↑	LLM-J ↑
		36.01	0.40	2.31
	✓	$69.07_{\uparrow 33.06}$	$0.45_{\uparrow 0.05}$	$4.21_{\uparrow 1.90}$
✓		$66.07_{\uparrow 30.06}$	$0.59_{\uparrow 0.19}$	$4.12_{\uparrow 1.81}$
✓	✓	$\mathbf{78.29}_{\uparrow 42.28}$	$\mathbf{0.67}_{\uparrow 0.27}$	$\mathbf{4.97}_{\uparrow 2.66}$

Table 2: **Ablation study** on Pri-Mixture.

Methods	Pri-Mixture ACC ↑	BS ↑	LLM-J ↑
DP-MLM$_{\text{oracle}}$	59.60	0.64	4.02
DP-MLM$_{\text{auto}}$	49.63	0.56	3.50
DAMPER$_{\text{oracle}}$	79.67	0.78	5.11
DAMPER	78.29	0.67	4.97

Table 3: **Robustness of localization** on Pri-Mixture under oracle vs. automatic (ours) localization.

hoc mechanisms. More results under varying privacy budgets are provided in Appendix G.1.

5.3 Diagnostic Analysis

Hyper-parameter Sensitivity. Figures 4a and 4b indicate that performance consistently peaks at $\beta = 0.1$ and $\tau_1 = 0.1$. Regarding the trade-off parameter α, Figure 4c confirms that lower values favor privacy (Semantic Obfuscation Index, SOI) while higher values prioritize fidelity (Domain Fidelity Score, DFS). To determine the optimal trade-off, we minimize the cumulative deviation from each metric's optimum (SUM-Drop, detailed in Appendix F.2). As shown in Figure 4d, $\alpha = 0.3$ achieves the most favorable balance. The optimal values of these parameters are consistent with the default settings used throughout our experiments.

Ablation Study. Table 2 presents the ablation results for $\mathcal{L}_{\text{CTR}}$ and $\mathcal{L}_{\text{DPO}}$ on the Pri-Mixture dataset. Individually, DPO and contrastive learning significantly boost ACC by 33.06% and 30.06%, respectively. The latter validates the necessity of mitigating embedding anisotropy, as discussed in Section 4.1. When combined, performance peaks at 42.28%, confirming the strong complementary effects of the two modules. Results for other datasets are detailed in Appendix G.3.

Robustness of Localization. Table 3 assesses the resilience of localization by benchmarking automatic inference (ours) against ground-truth oracle settings. DAMPER exhibits exceptional stability, maintaining performance comparable to DAMPER$_{\text{oracle}}$ (78.29% vs. 79.67%). In contrast, DP-MLM$_{\text{auto}}$ suffers a severe regression against DP-MLM$_{\text{oracle}}$ (49.63% vs. 59.60%), exposing a brittle dependency on ground-truth masks. More results of localization are provided in Appendix G.5.

We present results on **TextChunker's robustness** in Appendix G.4. We defer additional results on **single-domain training** to Appendix G.2, with extended **clustering analyses** available in Appendix G.6.

5.4 Generalization to Unseen Privacy Spans

Real-world privacy spans typically exhibit a long-tailed distribution (Zeng et al., 2025). To assess generalization to unseen or rare instances, we train DAMPER using only the Top-k% most frequent spans and evaluate it on the full test set. As illustrated in Figure 5, DAMPER displays exceptional robustness to vocabulary scarcity. It maintains 84.88% Recall even when restricted to the Top-20% frequent spans during training, with significant

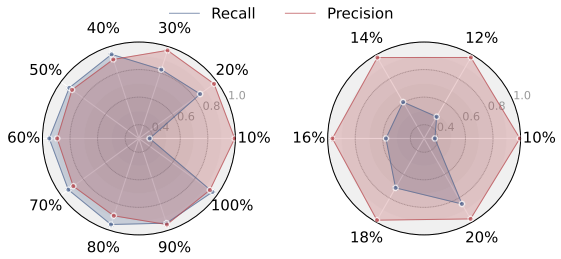


Figure 5: Performance under different **Top-%** predefined privacy spans on Pri-DDXPlus.

degradation appearing only below this threshold. This confirms that our learned prototypes capture abstract semantic regularities rather than relying on surface-level memorization, enabling effective detection extended to the long tail.

5.5 Privacy Attacks Assessment

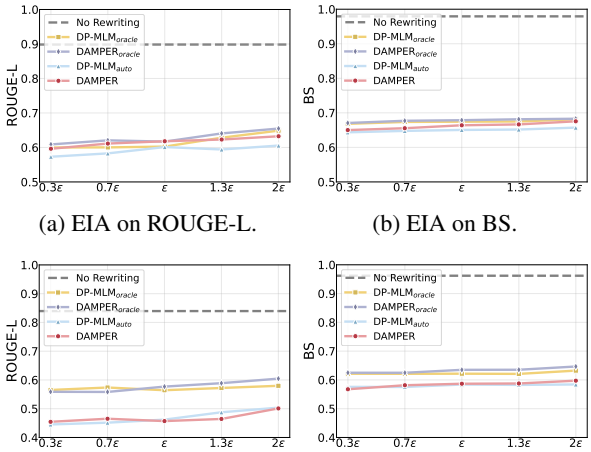
To rigorously evaluate the privacy robustness of DAMPER, we conduct two types of adversarial assessments against span-level rewriting methods following the methodology of Zeng et al. (2025).

Embedding Inversion Attack (EIA). EIA attempts to reconstruct sensitive spans directly from the user's input embeddings by training a dedicated inversion model. We quantify potential leakage using ROUGE-L (Lin, 2004) (see Appendix F.2) and BS measured between the reconstructed and ground-truth privacy spans. Empirical results (Figure 6a and 6b) indicate that across varying privacy budgets ϵ, DAMPER maintains a high level of protection comparable to other defense baselines.

Prompt Injection Attack (PIA). PIA aims to elicit the original input from the cloud LLM by injecting adversarial prompts into the rewritten text. We measure the attack success using ROUGE-L and BS between the recovered inputs and the original texts. Across all privacy budgets (Figure 6c and 6d), DAMPER and DP-MLM achieve similar defense capabilities under both oracle and automatic settings. The automatic setting consistently yields stronger privacy protection (i.e., lower leakage) than the oracle setting, as conservative detection often masks additional context. More details for both EIA and PIA are provided in Appendix G.7.

5.6 Computational Cost

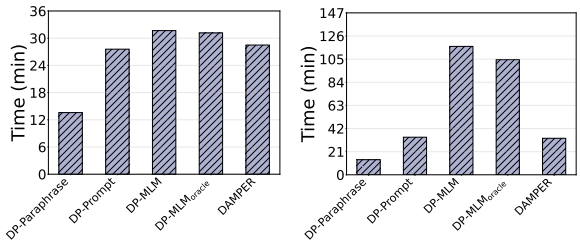
We evaluate DAMPER and four baselines, by performing privacy localization and rewriting on 1013 samples from Pri-DDXPlus and 509 samples from Pri-SLJA, respectively, and report the corresponding time costs. As illustrated in Figure 7, across both datasets, DP-MLM and its variants incur the highest time cost, while DP-Paraphrase is the most time-efficient. DAMPER exhibits nearly identical runtime to DP-Prompt. Notably, DP-MLM and its variants exhibit a pronounced increase in runtime as the input text length grows, whereas the other methods remain largely insensitive to text length. This behavior arises from the token-level rewriting strategy adopted by DP-MLM, which processes the input sequentially on a per-token basis.



(a) EIA on ROUGE-L. (b) EIA on BS.

(c) PIA on ROUGE-L. (d) PIA on BS.

Figure 6: Results of privacy attacks on span-level rewriting methods under varying ϵ on Pri-SLJA.



(a) Pri-DDXPlus. (b) Pri-SLJA.

Figure 7: **Computational cost comparison** of different methods on Pri-DDXPlus and Pri-SLJA

6 Conclusion

We proposed DAMPER, a client-side privacy rewriting framework for high-stakes, domain-sensitive queries under an untrusted cloud LLM. By inferring the dominant domain and localizing privacy spans with domain-contingent prototypes, DAMPER applies a calibrated DP mechanism only to the sensitive spans, thereby preserving non-sensitive context and task intent. Experiments on medical and legal benchmarks, including a challenging multi-domain setting, show that DAMPER improves utility and rewrite quality over strong baselines while reducing privacy leakage.

Acknowledgements

This work is partially supported by Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, the National Natural Science Foundation of China (NSFC Grant No.62406013, U25B2070, 62372493), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001), the Beijing Natural Science Foundation(Z230001).

Limitations

As an early step toward automatic privacy span localization for client-side rewriting, our work has several limitations. We evaluate only two domains and a constructed multi-domain setting, and future work should extend to a broader range of domains. Moreover, our current setting focuses on a multi-domain query stream, where each input is assumed to admit an identifiable dominant domain; a more challenging direction is single-query multi-domain fusion, in which cues from different domains are intertwined within the same input and may interfere with each other. We plan to investigate mechanisms to disentangle such fine-grained interactions in our future work.

Ethical Considerations

This research aligns strictly with the ACL Ethics Policy. Our experiments rely exclusively on Pri-DDXPlus, Pri-SLJA, and Pri-Mixture, which are derived from established open-source repositories and strictly follow the data protocols set by prior work (Zeng et al., 2025). We confirm that these benchmarks are fully anonymized and contain no real-world personally identifiable information (PII). Consequently, our study involves no interaction with human subjects, nor does it attempt to de-anonymize or reconstruct private data. It is important to note that our evaluation employs large language models (e.g., Qwen, GPT-4o). While these models facilitate automated rewriting and scoring, their outputs may reflect inherent biases or hallucinations common to pre-trained architectures.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023a. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023b. Frustratingly easy label projection for cross-lingual transfer. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pages 4299–4307.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

John C Duchi, Michael I Jordan, and Martin J Wainwright. 2013. Local privacy and statistical minimax rates. In *2013 IEEE 54th annual symposium on foundations of computer science*, pages 429–438. IEEE.

Cynthia Dwork. 2006. *Calibrating Noise to Sensitivity in Private Data Analysis*. Springer.

Cynthia Dwork and Aaron Roth. 2014. Algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Shuo Huang, Xingliang Yuan, Gholamreza Haffari, and Lizhen Qu. 2025. Zero-shot privacy-aware text rewriting via iterative tree search. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9175–9190, Suzhou, China. Association for Computational Linguistics.

Timour Igamberdiev, Thomas Arnold, and Ivan Habernal. 2022. DP-rewrite: Towards reproducibility and transparency in differentially private text rewriting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2927–2933, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: probing privacy leakage in large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14022–14040, Toronto, Canada. Association for Computational Linguistics.

Mingchen Li, Heng Fan, Song Fu, Junhua Ding, and Yunhe Feng. 2025. Dp-gtr: Differentially private prompt protection via group text rewriting. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1573–1585, Suzhou, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 867–881, Seattle, United States. Association for Computational Linguistics.

James B McQueen. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297.

Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103. 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, Bangkok, Thailand. Association for Computational Linguistics.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Zixuan Qin, Kunlin Lyu, Qingchen Yu, Yifan Sun, and Zhaoxin Fan. 2025. The achilles' heel of llms: How altering a handful of neurons can cripple language abilities. *Preprint*, arXiv:2510.10238.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Chelsea Finn, and Christopher Manning. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint*.

Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. 2019. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8934–8943.

Weiyan Shi, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. 2022. Just fine-tune twice: Selective differential privacy for large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6327–6340, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 4077–4087.

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

Xuchen Suo. 2024. Signed-prompt: A new approach to prevent prompt injection attacks against llm-integrated applications. In *AIP Conference Proceedings*, volume 3194, page 040013. AIP Publishing LLC.

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

Rushil Thareja, Preslav Nakov, Praneeth Vepakomma, and Nils Lukas. 2025. Dp-fusion: Token-level differentially private inference for large language models. *arXiv preprint arXiv:2507.04531*.

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, Singapore. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Qingchen Yu, Zifan Zheng, Ding Chen, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. GuessArena: Guess who I am? a self-adaptive framework for evaluating LLMs in domain-specific knowledge and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10897–10912, Vienna, Austria. Association for Computational Linguistics.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.

Ziqian Zeng, Jianwei Wang, Junyao Yang, Zhengdong Lu, Haoran Li, Huiping Zhuang, and Cen Chen. 2025. PrivacyRestore: Privacy-preserving inference in large language models via privacy removal and restoration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10821–10855, Vienna, Austria. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

Appendices

A Notations **12**

B Related Work **12**
- B.1 Token-level LDP Substitution . . . 12
- B.2 Sequence-level DP Paraphrasing . 12
- B.3 Span-level Privacy Rewriting . . . 12

C Preliminaries **14**
- C.1 Differential Privacy 14
- C.2 The Exponential Mechanism . . . 14
- C.3 Prototype Learning 14
- C.4 Direct Preference Optimization . . 15

D Details of TextChunker **15**

E Algorithms **15**

F More Experimental Setups **15**
- F.1 More Details of Datasets 15
- F.2 More Details of Metrics 16
- F.3 More Details of Implementation . 17

G Additional Experimental Results **18**
- G.1 More results of Performance Comparison 18
- G.2 Single-Domain Training 18
- G.3 More Results of Ablation Study . 18
- G.4 Results of TextChunker's Robustness 18
- G.5 More Results of Localization's Robustness 19
- G.6 Different Clustering Methods . . . 19
- G.7 Details of Privacy Attacks 20

H Qualitative Analysis **20**
- H.1 Case Studies 20
- H.2 Deployment Considerations 21

I Prompt Template Details **21**
- I.1 Candidate Generation 21
- I.2 Span-level Rewriting 21
- I.3 Accuracy Evaluation 21
- I.4 LLM-J Evaluation 21
- I.5 Prompt Injection Attack 21
- I.6 Prompt-Based Zero-Shot Localization 21

A Notations

Table 4 summarizes the notation used throughout this paper.

B Related Work

B.1 Token-level LDP Substitution

Local Differential Privacy (LDP) (Duchi et al., 2013) has been widely used to privatize text at the token level. Early mechanisms perturb each word independently and sample substitutes under metric-LDP constraints. SanText (Yue et al., 2021) selects replacements according to embedding distance to provide provable token-level guarantees, while CusText (Chen et al., 2023a) restricts the candidate neighborhood to reduce semantic drift. A key limitation is that token-level substitution effectively treats every token as sensitive, often leading to indiscriminate perturbation and stylistic or pragmatic mismatch.

B.2 Sequence-level DP Paraphrasing

To better exploit context, sequence-level approaches rewrite entire sentences or documents with DP mechanisms. DP-Paraphrase (Mattern et al., 2022) analyzes the limitations of token-level DP, highlighting unfavorable privacy–utility scaling with text length. DP-BART (Igamberdiev and Habernal, 2023) improves fluency with an encoder–decoder architecture and training-time DP techniques. DP-Prompt (Utpala et al., 2023) leverages LLM prompting for private rewriting without fine-tuning, and DP-MLM (Meisenbacher et al., 2024) proposes the 1-Diffractor mechanism to generate multiple DP-compliant candidates and select high-quality realizations. Recent work explores strategies that combine multiple granularities. DP-GTR (Li et al., 2025), for example, generates multiple DP paraphrases and identifies consensus keywords that consistently survive privatization for iterative suppression. Although these methods typically yield better coherence than token-level schemes, they still adopt a largely uniform notion of privacy across the input, hindering the tailoring of perturbations to regions with varying sensitivity.

B.3 Span-level Privacy Rewriting

Span-level methods focus on localizing and sanitizing sensitive segments while preserving the integrity of the surrounding non-sensitive context. Conventional approaches typically couple Named Entity Recognition (NER) or PII classifiers with selective sanitization, applying DP mechanisms strictly to tokens flagged as sensitive (Yue et al., 2021; Mattern et al., 2022). More advanced frameworks, such as PrivacyRestore (Zeng et al., 2025),

Symbol	Description
$\{a_i\}_{i=1}^{M}$, M	Spans generated by TextChunker$(\cdot)$ for input text x and their count
$\mathcal{D}$, $\mathcal{D}_{\text{pref}}$	Training corpus and preference dataset
$\mathcal{K}$, k	Total number of domains; domain index
$d_{k,i}$	Affinity score between span a_i and domain k, i.e., $d_{k,i} = \max_{1\le j\le J_k} \cos(\mathbf{z}_i, \mathbf{p}_{k,j})$
$\hat{k}$	Inferred domain index for input x
$score_i$	Privacy score for span a_i (used for thresholding)
γ_k	Privacy threshold for domain k
$\hat{\mathcal{S}}_{\hat{k}}^{x}$	Set of detected privacy spans in input x under inferred domain $\hat{k}$
$\mathcal{S}_k = \{s_{k,i}\}_{i=1}^{N_k}$, N_k	Annotated privacy spans and their count for domain k (training set)
$\mathcal{S}_k^x = \{s_{k,i}^x\}_{i=1}^{N_k^x}$, N_k^x	Annotated privacy spans in input x assigned to domain k and their count
$\mathcal{S}_k^y = \{s_{k,i}^y\}_{i=1}^{N_k^x}$	Corresponding set of rewritten spans aligned to $\mathcal{S}_k^x$
$\mathcal{U}$	Semantically meaningful spans
$h(\cdot)$, $g(\cdot)$	Trainable span encoder and frozen backbone encoder
$\mathbf{z}_{k,i}$	Representation of $s_{k,i}$ encoded by $h(\cdot)$, i.e., $\mathbf{z}_{k,i} = h(s_{k,i})$
$\mathbf{z}_{k,i}^x$, $\mathbf{z}_{k,i}^y$	Representations of $s_{k,i}^x$ and $s_{k,i}^y$ encoded by $h(\cdot)$
$\mathbf{z}_i$	Representation of span a_i encoded by $h(\cdot)$, i.e., $\mathbf{z}_i = h(a_i)$
$\mathcal{A}_{k,i}$	Positive set for anchor $s_{k,i}$ (in-domain privacy spans)
$\mathcal{N}_{k,i}$	Negative set for anchor $s_{k,i}$ (out-of-domain privacy spans and non-privacy spans)
$\mathcal{G}_{k,i}$	Union set $\mathcal{G}_{k,i} = \mathcal{A}_{k,i} \cup \mathcal{N}_{k,i}$ used in InfoNCE denominator
$\mathcal{L}_{CTR}$; $\mathcal{L}_{DPO}$	Multi-positive InfoNCE loss and DPO loss
τ_1	Temperature for contrastive loss
$\mathcal{P}_k = \{\mathbf{p}_{k,j}\}_{j=1}^{J_k}$, J_k	Set of privacy prototypes and the number of prototypes for domain k
$j_i^\star$	Index of the nearest prototype for the i-th rewritten span, $j_i^\star = \arg\max_{1\le j\le J_k} \cos(\mathbf{z}_{k,i}^y, \mathbf{p}_{k,j})$
$\mathcal{Y}(x) = \{y^{(c)}\}_{c=1}^{C}$	Set of candidate rewrites for input x
y_w, y_l	Preferred rewrite (winner) and dispreferred rewrite (loser)
$r_{\text{priv}}(\cdot)$, $r_{\text{util}}(\cdot)$, $r(\cdot)$	Privacy reward, utility reward, and composite reward
α	Weighting parameter balancing privacy and utility rewards
β	Hyperparameter for DPO
π_θ, π_{ref}	Rewriter policy and reference policy
x; y	Input text and rewritten text
$\mathcal{V}$	Vocabulary of the rewriter π_θ
u_ℓ, $\bar{u}_\ell$	Logit vector at decoding step ℓ and its clipped version, $\bar{u}_\ell = \text{clip}(u_\ell; R_1, R_2)$
R_1, R_2	Lower bound and upper bound for logit clipping
Δu	Utility sensitivity
τ_2	Sampling temperature for the rewriter (DP mechanism)
ϵ_{token}, ϵ_{text}, ϵ	Token-level privacy cost and text-level (query-level) privacy budget, privacy hyperparameter
$n_{\text{sp}}(x)$, $n_{\text{sp}}^{\max}$	Number of regenerated tokens within detected spans in x; global upper bound used for budgeting
$\sigma(\cdot)$	Sigmoid function

Table 4: Summary of notation used in this paper.

decompose the process into client-side removal and server-side restoration to enhance downstream utility. Recently, NaPaRe (Huang et al., 2025) introduced a zero-shot framework that utilizes iterative tree search for controllable rewriting (e.g., abstraction) without task-specific fine-tuning. Similarly, DP-Fusion (Thareja et al., 2025) proposes a differentially private inference mechanism that mixes output distributions from original and redacted contexts to provably bound the influence of identified sensitive token groups.

Despite these advancements in locality and controllability, existing pipelines face critical barriers to deployable client-side use. A substantial limitation lies in the rigidity of detection mechanisms. These methods inherently assume that sensitive spans can be exhaustively identified via static inventories or explicit user cues—which precludes autonomous localization and severely restricts adaptability to domain-contingent semantics. Beyond detection, distinct theoretical and architectural constraints persist. Recent search-based systems often prioritize empirical robustness against reconstruction attacks over rigorous formalism, lacking the formal span-level DP guarantees necessary for worst-case protection. Furthermore, frameworks like PrivacyRestore necessitate restoration mechanisms that are tightly coupled to specific server-side LLMs, preventing their deployment in model-agnostic scenarios where the downstream service is an opaque black-box API.

In contrast, our framework integrates domain-

specific representation learning with prototype-guided localization and a DP-compliant rewriting mechanism. This design achieves autonomous, domain-contingent sanitization with formal privacy guarantees, while remaining fully decoupled from downstream LLM services.

C Preliminaries

C.1 Differential Privacy

Differential Privacy (DP) (Dwork, 2006; Dwork and Roth, 2014) protects the contribution of any single individual in a dataset. Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the output space. Two inputs $x, x' \in \mathcal{X}$ are called *neighbors*, written $x \sim x'$, if they differ in the contribution of one individual (or, in the local setting, correspond to two possible private values of a single user). A randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is said to satisfy ϵ-DP if for all $x \sim x'$ and measurable $S \subseteq \mathcal{Y}$,

$$\Pr[\mathcal{M}(x) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(x') \in S]. \tag{15}$$

where $\epsilon > 0$ is the privacy budget. In the context of text rewriting, we often operate under Local Differential Privacy (LDP) (Duchi et al., 2013), where the sanitization mechanism is applied locally by the user before data leaves their device.

C.2 The Exponential Mechanism

The Exponential Mechanism (EM) (McSherry and Talwar, 2007) is a fundamental technique for selecting outputs from a discrete domain $\mathcal{Y}$ while preserving differential privacy. Given an input x, a utility function $u : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ assigns a real-valued score to each possible output $y \in \mathcal{Y}$, reflecting its quality or relevance. The privacy guarantee of the mechanism depends on the global sensitivity of the utility function, defined as the maximum change in utility caused by modifying a single input entry:

$$\Delta u = \max_{x \sim x'} \max_{y \in \mathcal{Y}} |u(x, y) - u(x', y)|. \tag{16}$$

The Exponential Mechanism $\mathcal{M}_{\text{EM}}(x)$ selects an output y with probability proportional to its utility, scaled by the privacy budget ϵ and sensitivity Δu:

$$\Pr[\mathcal{M}_{\text{EM}}(x) = y] = \frac{\exp\left(\frac{\epsilon \cdot u(x,y)}{2\Delta u}\right)}{\sum_{y' \in \mathcal{Y}} \exp\left(\frac{\epsilon \cdot u(x,y')}{2\Delta u}\right)}. \tag{17}$$

This distribution ensures that outputs with higher utility are exponentially more likely to be chosen, providing a theoretically grounded trade-off between utility and privacy.

Sampling-based Exponential Mechanism. In the context of autoregressive or masked language generation, the output domain corresponds to the vocabulary $\mathcal{V}$ at each decoding step. Recent works operationalize the Exponential Mechanism directly via the model's standard sampling process. Specifically, the unnormalized log-probability (logit) $v_t \in \mathbb{R}$ for a token $t \in \mathcal{V}$ serves as the utility function, i.e., $u(x, t) = v_t$. The standard softmax sampling distribution with a temperature $\tau > 0$ is given by:

$$\Pr(t|x) = \frac{\exp(v_t/\tau)}{\sum_{t' \in \mathcal{V}} \exp(v_{t'}/\tau)}. \tag{18}$$

By equating the exponents in Eq. (17) and the softmax function, we observe that setting the temperature $\tau = \frac{2\Delta u}{\epsilon}$ renders the sampling process formally equivalent to the Exponential Mechanism (Mattern et al., 2022; Utpala et al., 2023). However, the raw logits of neural language models are theoretically unbounded, implying an infinite sensitivity Δu. To satisfy DP guarantees, it is necessary to bound the utility function. This is typically achieved by clipping the logits to a fixed range $[R_1, R_2]$ (or applying a norm bound) before sampling. Under this constraint, the sensitivity is upper-bounded by $\Delta u \leq 2(R_2 - R_1)$ (the maximum distance between any two values in $[R_1, R_2]$). Thus, sampling from the clipped logits with temperature τ satisfies ϵ-DP at the token level, where:

$$\epsilon = \frac{4(R_2 - R_1)}{\tau}. \tag{19}$$

This formulation allows standard LLMs to function as differentially private generators by simply adjusting the sampling temperature and enforcing logit constraints.

C.3 Prototype Learning

Prototype learning maps inputs into a latent embedding space where classes or clusters are represented by identifying *prototypes*. In supervised settings, such as Prototypical Networks (Snell et al., 2017), prototypes are typically computed as the mean embedding of support examples for each class. In unsupervised or semi-supervised settings, prototypes can be derived via clustering algorithms (e.g., k-means (McQueen, 1967), GMM (Dempster et al., 1977) or FINCH (Sarfraz et al., 2019)) to serve as structured, discrete summaries of the underlying data distribution.

C.4 Direct Preference Optimization

Preference alignment aims to steer language models toward desired behaviors using relative feedback rather than absolute rewards (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022). Direct Preference Optimization (DPO) (Rafailov et al., 2023) optimizes a policy π_θ directly from a dataset of preference pairs (x, y_w, y_l), where y_w is preferred over y_l. Unlike Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), which requires training a separate reward model, DPO derives a closed-form objective by implicitly defining the reward via the ratio of the policy likelihood to a frozen reference model π_{ref}. The objective minimizes the negative log-likelihood of the preference data:

$$-\mathbb{E}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}-\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\Big]. \tag{20}$$

where β is a hyperparameter controlling the deviation from the reference policy. This approach provides a stable and computationally efficient method for aligning model outputs with complex preference criteria.

D Details of TextChunker

We adopt a rule-based and syntactically guided text chunking approach to decompose the input text into semantically coherent spans. Given an input text, we first perform sentence segmentation and shallow syntactic analysis using spaCy with the `en_core_web_sm` model. These annotations provide sufficient structural information for phrase-level reasoning without requiring full semantic parsing. Inspired by Lin et al. (2020), for each sentence, we then identify a small set of semantic frame triggers, implemented as regular-expression patterns. Once a trigger is detected, the text following the trigger is treated as a candidate enumeration region and is split using weak separators, including commas, semicolons, and coordinating conjunctions (*and*, *or*). This design enables robust handling of parallel structures while avoiding overly aggressive segmentation. To improve recall under more complex or implicit sentence constructions, we additionally incorporate syntax-aware phrase extraction. Concretely, we collect noun phrases identified by spaCy's noun chunker, gerundive verb phrases whose head tokens are tagged as VBG (Marcus et al., 1993) (with the span defined by the corresponding dependency subtree), and infinitive purpose constructions matching the pattern *to verb*. These phrase types are widely recognized as minimal semantic carriers in span-based modeling and information extraction (Joshi et al., 2020; Chen et al., 2023b). Finally, we prevent spans from crossing sentence-internal strong punctuation and normalize span boundaries by removing leading conjunctions and trailing punctuation.

Algorithm 1 Learning Domain Privacy Prototypes

Require: Annotated privacy spans $\{\mathcal{S}_k\}$, backbone $g(\cdot)$, temperature τ_1.
Ensure: Privacy prototypes $\{\mathcal{P}_k\}_{k\in\mathcal{K}}$

```
Initialize span encoder h(·) from g(·).
// Contrastive Representation Learning
for each training step do
    Sample batch of anchor spans s_{k,i};
    Get positive set A_{k,i} and negative set G_{k,i};
    Update h(·) by minimizing L_CTR (Eq. (1)).
end for
// Prototype Clustering
for each domain k ∈ K do
    Extract z_k = {h(s) | s ∈ S_k}.
    Cluster P_k ← FINCH(z_k).
end for
return {P_k}
```

E Algorithms

The algorithmic details of DAMPER are presented below. Algorithm 3 outlines the Offline Training Phase, coordinating prototype learning (Algorithm 1) and preference construction (Algorithm 2). The Online Inference Phase is detailed in Algorithm 4.

F More Experimental Setups

F.1 More Details of Datasets

We evaluate our framework on two established domain-specific datasets and a synthesized multi-domain benchmark constructed as follows.

Pri-DDXPlus. This dataset focuses on the medical diagnosis domain (Zeng et al., 2025). Each sample consists of a patient symptom description, explicitly annotated with privacy-sensitive and non-sensitive spans based on medical confidentiality standards. The task is formulated as a multiple-choice question, where each sample is associated with one correct diagnosis and three randomly sampled incorrect diagnoses as in-domain distractors.

Pri-SLJA. Analogous to Pri-DDXPlus, this dataset

Algorithm 2 Preference Dataset Construction

Require: Training corpus $\mathcal{D}$, reference model π_{ref}, prototypes $\{\mathcal{P}_k\}$, span encoder $h(\cdot)$, hyperparameter α.
Ensure: Preference dataset $\mathcal{D}_{\text{pref}}$.

Initialize $\mathcal{D}_{\text{pref}} \leftarrow \emptyset$.
for each input x in $\mathcal{D}$ **do**
  $\mathcal{Y}(x) \leftarrow \{y^{(n)} \sim \pi_{\text{ref}}(x, \mathcal{S}_k^x)\}_{n=1}^N$.
  // Composite Reward Scoring
  for each candidate $y \in \mathcal{Y}(x)$ **do**
    Compute $r_{\text{priv}}(y)$ via Eq. (3);
    Compute $r_{\text{util}}(y)$ via Eq. (4);
    $r(y) \leftarrow (1-\alpha) r_{\text{priv}}(y) + \alpha r_{\text{util}}(y)$.
  end for
  Select $y_w \leftarrow \arg\max_{y \in \mathcal{Y}(x)} r(y)$;
  Select $y_l \leftarrow \arg\min_{y \in \mathcal{Y}(x)} r(y)$;
  $\mathcal{D}_{\text{pref}} \leftarrow \mathcal{D}_{\text{pref}} \cup \{(x, y_w, y_l)\}$.
end for
return $\mathcal{D}_{\text{pref}}$

Algorithm 3 DAMPER Offline Training Phase

Require: Training corpus $\mathcal{D}$, annotated spans $\{\mathcal{S}_k\}$, backbone $g(\cdot)$, reference model π_{ref}.
Ensure: Optimized rewriter π_θ.

// Prototype Learning
$\{\mathcal{P}_k\}, h(\cdot) \leftarrow$ Algorithm 1$(\{\mathcal{S}_k\}, g)$
// Preference Construction
$\mathcal{D}_{\text{pref}} \leftarrow$ Algorithm 2$(\mathcal{D}, \pi_{\text{ref}}, \{\mathcal{P}_k\}, h)$
// Preference Alignment
Initialize π_θ from π_{ref}.
while not converged **do**
  Sample batch (x, y_w, y_l) from $\mathcal{D}_{\text{pref}}$.
  Update π_θ by minimizing $\mathcal{L}_{\text{DPO}}$ (Eq. (6)).
end while
return π_θ

targets the legal judgment domain (Zeng et al., 2025). Each entry contains a detailed description of a legal case, with span-level annotations distinguishing privacy-sensitive details from general context according to legal norms. Similarly, each sample is paired with one correct judgment and three random incorrect judgments.

Pri-Mixture. To assess performance in multi-domain scenarios, we construct Pri-Mixture by augmenting the two base datasets with multi-domain distractors. Specifically, we append four law-related options to the original medical samples (Pri-DDXPlus) and four medical-related options to the legal samples (Pri-SLJA). These augmented datasets are then amalgamated to form a unified

Algorithm 4 Online Inference Phase

Require: User query x, span encoder $h(\cdot)$, prototypes $\{\mathcal{P}_k\}$, rewriter π_θ.
Ensure: rewritten query y.

// Text Segmentation
Segment $\{a_i\}_{i=1}^M \leftarrow$ TEXTCHUNKER(x);
Compute $\mathbf{z}_i \leftarrow h(a_i)$ for $i = 1 \dots M$.
// Domain Inference
for each domain $k \in \mathcal{K}$ **do**
  for each span $i = 1 \dots M$ **do**
    Calculate $d_{k,i} \leftarrow \max_{\mathbf{p} \in \mathcal{P}_k} \cos(\mathbf{z}_i, \mathbf{p})$.
  end for
  Domain score: $s_k \leftarrow \frac{1}{M} \sum_{i=1}^M d_{k,i}$.
end for
Infer global domain: $\hat{k} \leftarrow \arg\max_{k \in \mathcal{K}} s_k$.
// Privacy Span Localization
Initialize privacy set $\hat{\mathcal{S}}^x \leftarrow \emptyset$.
sort affinities: $d_{\hat{k},(1)} \leq d_{\hat{k},(2)} \leq \cdots \leq d_{\hat{k},(M)}$
for $t = 1 \dots M$ **do**
  Calculate $\sigma^2(t)$ by Eq. (9).
end for
Find $t^\star \leftarrow \arg\max_t \sigma^2(t)$.
Threshold: $\gamma_{\hat{k}} \leftarrow (d_{\hat{k},(t^\star)} + d_{\hat{k},(t^{\star+1})})/2$
for each span $i = 1 \dots M$ **do**
  if $d_{\hat{k},i} > \gamma_{\hat{k}}$ **then**
    $\hat{\mathcal{S}}^x \leftarrow \hat{\mathcal{S}}^x \cup \{a_i\}$.
  end if
end for
// DP Span-level Rewriting
Apply rewriter only on detected spans:
$\tilde{y} = \tilde{\pi}_\theta(x, \hat{\mathcal{S}}_{\hat{k}}^x)$ with EM (Eq. (11)).

multi-domain benchmark, challenging the model to distinguish correct answers amidst both in-domain and out-of-domain noise.

F.2 More Details of Metrics

We employ a comprehensive suite of metrics to evaluate our framework across three critical dimensions: downstream utility, semantic fidelity, and privacy preservation.

Accuracy (Acc ↑). To assess the utility of the rewritten text for downstream applications, we measure the prediction accuracy of a cloud-based LLM on the processed queries. This evaluation is standardized using a structured prompt template (see Appendix I.3) to ensure consistent decision-making logic across different test cases.

BERTScore (BS ↑). BERTScore quantifies the semantic similarity between the original query x and the rewritten output y. By computing the co-

sine similarity of their contextual embeddings via a pre-trained encoder, it serves as a robust proxy for measuring how well the semantic content is preserved after rewriting.

LLM-Judge (LLM-J ↑). Following recent benchmarks (Zheng et al., 2023; Zeng et al., 2025), we employ an LLM-Judge to evaluate the overall quality and coherence of the generated text. Specifically, we first generate responses using Qwen2.5-7B-Instruct (Team, 2024; Yang et al., 2024) based on the rewritten queries, and then leverage DeepSeek-V3.2 (DeepSeek-AI, 2025) as an evaluator to assign a quality score on a scale of 1 to 10. The prompt templates used for this evaluation are provided in Appendix I.4.

Semantic Obfuscation Index (SOI ↑) and Domain Fidelity Score (DFS ↑). These dual metrics quantify the effectiveness of the rewriting policy in balancing competing objectives. **SOI** measures the degree of privacy masking, directly corresponding to the privacy reward defined in Eq. (3), while **DFS** assesses the preservation of domain-specific characteristics, aligning with the utility reward in Eq. (4). To ensure a focused assessment, both metrics are computed exclusively on the localized privacy spans rather than the entire text.

SOI-Drop (↓) and DFS-Drop (↓). To analyze the sensitivity of model performance to the trade-off parameter α, we introduce the Drop metrics. Let $\mathrm{M} \in \{\mathrm{SOI}, \mathrm{DFS}\}$ denote the min-max normalized score for a given metric. The performance drop is defined as the deviation from the optimal achievable value:

$$\text{M-Drop} = \frac{\max(\mathrm{M}) - \mathrm{M}}{\max(\mathrm{M})}. \tag{21}$$

This metric helps in identifying the parameter configuration that minimizes the collective loss in privacy and utility.

Privacy F1-Score (PF1 ↑). To provide a holistic assessment of the span localization module, we employ the F1-Score, which balances the trade-off between Precision and Recall. Precision quantifies the proportion of predicted spans that are actual ground-truth privacy spans:

$$\text{Precision} = \frac{|S_k^x \cap \hat{S}_{\hat{k}}^x|}{|\hat{S}_{\hat{k}}^x|}, \tag{22}$$

while Recall measures the proportion of ground-truth spans that are successfully retrieved by the model:

$$\text{Recall} = \frac{|S_k^x \cap \hat{S}_{\hat{k}}^x|}{|S_k^x|}. \tag{23}$$

The unified PF1 metric is calculated as the harmonic mean:

$$\text{PF1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{24}$$

ROUGE-L (↓). ROUGE-L (Lin, 2004) measures the structural similarity between the generated sequence and the reference sequence based on the longest common subsequence. In our privacy context, we calculate this metric between the original sensitive spans and their rewritten counterparts. A lower ROUGE-L score indicates less lexical overlap, reflecting more effective obfuscation of sensitive information.

F.3 More Details of Implementation

DAMPER Training Configurations. The span encoder is fine-tuned on 149 medical and 142 legal privacy span types. We apply LoRA with a rank of 16 and train for 30 epochs using a batch size of 32 and the AdamW optimizer (learning rate 10^{-4}). For preference construction and DPO, we utilize LLaMA-Factory (Zheng et al., 2024) to generate 10 candidates per sample. The policy model is fine-tuned for 7 epochs with LoRA (rank 16), a batch size of 16, and a learning rate of 5×10^{-6}. Regarding dataset-specific hyperparameters, we set $R_1 = 5$ and $R_2 = 20$.

Baseline Configurations. To ensure a fair comparison, all baselines are evaluated under the same total privacy budget constraint (ϵ_{text}) as DAMPER. For **DP-Paraphrase**, we adopt GPT-2, fine-tuned by Meisenbacher et al. (2024), as the paraphrasing model. For **DP-Prompt**, we utilize the open-source Flan-T5-xl (Chung et al., 2022) to balance performance and adaptability. Regarding **DP-MLM** and **DP-MLM$_{\text{oracle}}$**, we employ RoBERTa-base (Liu et al., 2019), concatenating the original text as auxiliary guidance during inference. Finally, for **PrivacyRestore**, to maintain comparability with DAMPER, we use Qwen2.5-7B-Instruct as the inference model and BERT-base-uncased (Devlin et al., 2019) as the auxiliary retrieval model, with the number of common attention heads set to 175. All remaining configurations also follow the default settings reported in their original papers.

Privacy Budget Allocation. Following PrivacyRestore (Zeng et al., 2025), we fix privacy hyperparameter ϵ to 75 and report ϵ_{text} and ϵ_{token} under ϵ.

Datasets	Full-Text Level		Span Level (oracle)		Span Level (auto)		ϵ_{text}
	n^{max}	ϵ_{token}	$n_{\text{sp}}^{\text{max}}$	ϵ_{token}	$n_{\text{sp}}^{\text{max}}$	ϵ_{token}	
Pri-DDXPlus	86	1.74	49	3.06	52	2.88	150
Pri-SLJA	178	0.84	41	3.66	104	1.44	150

Table 5: Privacy budget settings of different baselines across Pri-DDXPlus and Pri-SLJA. Baselines are categorized into three levels: **Full-Text Level**, including DP-Paraphrase, DP-Prompt, and DP-MLM; **Span Level (oracle)**, including DP-MLM$_{\text{oracle}}$ and DAMPER$_{\text{oracle}}$; and **Span Level (auto)**, including DP-MLM$_{\text{auto}}$ and DAMPER.

Table 5 provides a detailed breakdown of the privacy budget allocation under the setting $\epsilon_{\text{text}} = 2\epsilon$. Since baselines operate at different granularities (Full-text vs. Span-level), we convert the total budget ϵ_{text} into equivalent per-token budgets ϵ_{token} based on the maximum sequence lengths n^{max} or maximum span lengths $n_{\text{sp}}^{\text{max}}$.

All experiments were conducted under Python 3.12 on a system equipped with two **NVIDIA RTX A6000 GPUs** (48 GB each).

G Additional Experimental Results

G.1 More results of Performance Comparison

We further conduct additional experiments under different privacy budget configurations. Specifically, the ϵ_{token} of DAMPER is set to 9.62 on Pri-DDXPlus and 4.81 on Pri-SLJA. The setting of ϵ_{text} and the ϵ_{token} for the other baselines follow the same protocol as in Table 5. The results are reported in Table 6.

G.2 Single-Domain Training

We use the single dataset during training phase. As shown in Table 7, we compare the evaluation results of single-domain training with multi-domain training of our method and other baselines. The performance on the two single-domain datasets is nearly comparable to that achieved with multi-domain training. On Pri-DDXPlus, a slight degradation is observed in both rewriting quality and task accuracy, whereas on Pri-SLJA, single-domain training yields marginally better performance. These results indicate that DAMPER can effectively adapt to multi-domain rewriting tasks.

G.3 More Results of Ablation Study

In this section, we provide comprehensive ablation results for the individual datasets, **Pri-DDXPlus** and **Pri-SLJA**, which align with the mixture dataset observations presented in the main text (Table 8). On **Pri-DDXPlus**, the combination of $\mathcal{L}_{CTR}$ and $\mathcal{L}_{DPO}$ yields an ACC improvement of 44.02 % over the baseline, surpassing the gains from using either module individually. Consistent with this trend, the combined approach on **Pri-SLJA** achieves the highest performance across all metrics, reaching an ACC of 82.68% and a BS of 0.66. These results align with the analysis on Pri-Mixture, robustly validating the effectiveness and necessity of each component in our DAMPER.

G.4 Results of TextChunker's Robustness

We conducted a robustness study on segmentation by replacing the original TextChunker with alternative span proposal strategies. Specifically, we considered the following variants. **Boundary Perturbation (BP)**: Starting from the original TextChunker outputs, we randomly select a proportion p (in our setting, $p = 0.2$) of spans and apply random re-segmentation. **YAKE-based Segmentation (YAKE)**(Campos et al., 2020): We use YAKE, an unsupervised keyphrase extraction method that scores candidate phrases based on local statistical features such as term frequency, positional information, and contextual distribution. The extracted keyphrases are treated as candidate spans. **n-gram Enumeration (n-gram)**: We enumerate all contiguous n-grams with length $L \leq 6$ as candidate spans. This strategy increases coverage while introducing denser span proposals. The experimental results in Table 9 show that, on both Pri-DDXPlus and Pri-SLJA, DAMPER achieves comparable ACC and PF1 across all four span proposal strategies. The differences between the original TextChunker and the alternative segmentation methods are marginal, indicating that the overall privacy–utility trade-off is not strongly dependent on a specific segmentation heuristic. These findings suggest that the currently adopted TextChunker provides stable and effective span granularity for prototype-based localization. At the same time, since the segmentation step only defines candidate spans and does not determine sensitivity, the framework remains compatible with alternative or more advanced span proposal mechanisms.

Methods	Pri-DDXPlus			Pri-SLJA			Pri-Mixture		
	ACC ↑	BS ↑	LLM-J ↑	ACC ↑	BS ↑	LLM-J ↑	ACC ↑	BS ↑	LLM-J ↑
No Rewriting	$87.60_{\pm 0.07}$	$1.00_{\pm 0.00}$	$5.60_{\pm 0.14}$	$89.79_{\pm 0.28}$	$1.00_{\pm 0.00}$	$6.89_{\pm 0.08}$	$88.45_{\pm 0.14}$	$1.00_{\pm 0.00}$	$6.04_{\pm 0.12}$
DP-Paraphrase	$46.20_{\pm 0.08}$	$0.44_{\pm 0.00}$	$1.74_{\pm 0.01}$	$47.45_{\pm 0.13}$	$0.41_{\pm 0.00}$	$1.95_{\pm 0.07}$	$43.49_{\pm 0.19}$	$0.43_{\pm 0.00}$	$1.81_{\pm 0.02}$
DP-Prompt	$30.37_{\pm 1.29}$	$0.29_{\pm 0.00}$	$1.72_{\pm 0.08}$	$25.15_{\pm 0.83}$	$0.23_{\pm 0.00}$	$2.24_{\pm 0.13}$	$27.16_{\pm 1.09}$	$0.27_{\pm 0.00}$	$1.90_{\pm 0.01}$
DP-MLM	$34.40_{\pm 0.96}$	$0.40_{\pm 0.03}$	$1.86_{\pm 0.04}$	$29.81_{\pm 0.42}$	$0.29_{\pm 0.08}$	$2.41_{\pm 0.16}$	$31.16_{\pm 0.28}$	$0.35_{\pm 0.04}$	$2.05_{\pm 0.02}$
DP-MLM$_{\text{oracle}}$	$52.19_{\pm 1.77}$	$\underline{0.60}_{\pm 0.03}$	$3.87_{\pm 0.08}$	$\underline{81.14}_{\pm 1.67}$	$\mathbf{0.75}_{\pm 0.01}$	$5.07_{\pm 0.16}$	$61.27_{\pm 1.44}$	$\underline{0.66}_{\pm 0.01}$	$4.28_{\pm 0.00}$
PrivacyRestore	$\underline{76.79}_{\pm 0.85}$	-	$\underline{4.41}_{\pm 0.13}$	$79.36_{\pm 0.54}$	-	$\mathbf{5.62}_{\pm 0.04}$	$\underline{77.32}_{\pm 0.64}$	-	$\underline{4.82}_{\pm 0.07}$
DAMPER	$\mathbf{79.71}_{\pm 0.03}$	$\mathbf{0.71}_{\pm 0.00}$	$\mathbf{4.96}_{\pm 0.05}$	$\mathbf{83.17}_{\pm 0.78}$	$\underline{0.68}_{\pm 0.00}$	$\underline{5.53}_{\pm 0.02}$	$\mathbf{80.01}_{\pm 0.41}$	$\mathbf{0.69}_{\pm 0.00}$	$\mathbf{5.15}_{\pm 0.02}$

Table 6: More results of performance comparison on three datasets. ACC is reported in %, where "-" denotes unavailable results. Results are averaged over 3 runs. The **best** and second are marked.

Methods	Pri-DDXPlus		
	ACC ↑	BS ↑	LLM-J ↑
DP-MLM$_{\text{oracle}}$	50.94	0.59	3.61
PrivacyRestore	75.82	-	4.32
DAMPER$_{\text{multi}}$	**78.13**	**0.68**	**4.76**
DAMPER$_{\text{single}}$	77.83	0.61	4.69
Methods	**Pri-SLJA**		
	ACC ↑	BS ↑	LLM-J ↑
DP-MLM$_{\text{oracle}}$	78.38	**0.73**	4.81
PrivacyRestore	78.38	-	5.26
DAMPER$_{\text{multi}}$	82.68	0.66	5.37
DAMPER$_{\text{single}}$	**82.71**	0.55	**5.66**

Table 7: Performance comparison on the single-domain. **DAMPER$_{\text{multi}}$** denotes the setting where both Pri-DDXPlus and Pri-SLJA are jointly used during training (evaluated on the corresponding dataset), while **DAMPER$_{\text{single}}$** refers to training using only a single dataset.

$\mathcal{L}_{CTR}$	$\mathcal{L}_{DPO}$	Pri-DDXPlus ACC ↑	BS ↑	LLM-J ↑
		34.11	0.42	2.22
	✓	$66.50_{\uparrow 32.39}$	$0.48_{\uparrow 0.06}$	$3.83_{\uparrow 1.61}$
✓		$65.59_{\uparrow 31.48}$	$0.61_{\uparrow 0.19}$	$3.91_{\uparrow 1.69}$
✓	✓	$\mathbf{78.13}_{\uparrow 44.02}$	$\mathbf{0.68}_{\uparrow 0.26}$	$\mathbf{4.76}_{\uparrow 2.54}$
$\mathcal{L}_{CTR}$	$\mathcal{L}_{DPO}$	**Pri-SLJA** ACC ↑	BS ↑	LLM-J ↑
		40.08	0.36	2.47
	✓	$74.46_{\uparrow 34.38}$	$0.39_{\uparrow 0.03}$	$4.94_{\uparrow 2.47}$
✓		$69.49_{\uparrow 29.41}$	$0.56_{\uparrow 0.20}$	$4.53_{\uparrow 2.06}$
✓	✓	$\mathbf{82.68}_{\uparrow 42.60}$	$\mathbf{0.66}_{\uparrow 0.30}$	$\mathbf{5.37}_{\uparrow 2.90}$

Table 8: **Ablation study** on the performance of $\mathcal{L}_{CTR}$ and $\mathcal{L}_{DPO}$. The **best** are marked.

G.5 More Results of Localization's Robustness

Table 10 shows DAMPER demonstrates remarkable resilience: on Pri-DDXPlus, accuracy remains stable, shifting marginally from 78.12% to 78.13%, while on the challenging Pri-SLJA domain, the performance decline is well-contained from 85.27% to 82.68%. In sharp contrast, DP-MLM suffers catastrophic degradation on Pri-SLJA, with accuracy plummeting from 78.38% to 55.76%, revealing a critical dependency on perfect masks. These results

Methods	Pri-DDXPlus		Pri-SLJA	
	ACC ↑	PF1 ↑	ACC ↑	PF1 ↑
BP	77.53	94.84	82.12	67.15
YAKE	79.15	95.75	79.76	68.69
n-gram	78.24	85.18	78.98	58.41
Original	78.13	94.84	82.68	67.15

Table 9: Performance comparison across different TextChunker variants.

Methods	Pri-DDXPlus		
	ACC ↑	BS ↑	LLM-J ↑
DP-MLM$_{\text{oracle}}$	50.94	0.59	3.61
DP-MLM$_{\text{auto}}$	48.89	0.56	3.44
DAMPER$_{\text{oracle}}$	78.12	0.72	4.72
DAMPER	78.13	0.68	4.76
Methods	**Pri-SLJA**		
	ACC ↑	BS ↑	LLM-J ↑
DP-MLM$_{\text{oracle}}$	78.38	0.73	4.81
DP-MLM$_{\text{auto}}$	55.76	0.57	3.61
DAMPER$_{\text{oracle}}$	85.27	0.90	5.86
DAMPER	82.68	0.66	5.37

Table 10: Performance comparison of different module combinations.

confirm that our prototype-guided alignment significantly fortifies end-to-end performance against detection noise, offering a far more robust solution than generic masking baselines.

G.6 Different Clustering Methods

Table 11 compares the performance of three clustering algorithms. The Mean method constructs a domain prototype by averaging all privacy embeddings within each domain, while k-means applies clustering with the number of clusters set to 4 for each domain. FINCH almost achieves the best performance across all evaluation metrics on all three datasets. In contrast, the Mean method assigns only a single prototype to each domain, which limits its ability to adequately cover the diversity of domain-specific private spans, leading to degraded overall utility. Furthermore, k-means implicitly assumes that clusters exhibit near-spherical structures in the embedding space. However, this assumption

Method	#Protos	Ratio	Pri-DDXPlus		
			ACC ↑	BS ↑	LLM-J ↑
Mean	1	0.007	73.79	0.67	4.29
k-means	4	0.027	73.68	0.67	4.31
FINCH	3	0.020	**78.13**	**0.68**	**4.76**
Method	#Protos	Ratio	Pri-SLJA		
			ACC ↑	BS ↑	LLM-J ↑
Mean	1	0.007	78.59	0.64	5.36
k-means	4	0.028	78.78	0.64	**5.38**
FINCH	4	0.028	**82.68**	**0.66**	5.37
Method	#Protos	Ratio	Pri-Mixture		
			ACC ↑	BS ↑	LLM-J ↑
Mean	2	0.007	75.35	0.66	4.65
k-means	8	0.027	75.75	0.66	4.67
FINCH	7	0.024	**78.29**	**0.67**	**4.97**

Table 11: Comparison of different clustering algorithms on three datasets. **#Protos** denotes the number of prototypes and **Ratio** is defined as the ratio between the number of prototypes and the total number of privacy span types. The **best** are marked.

is frequently violated in highly anisotropic high-dimensional representations, leading to suboptimal performance relative to the FINCH.

G.7 Details of Privacy Attacks

To evaluate the privacy protection capability of DAMPER, we conduct adversarial attacks under a realistic threat model. Specifically, the adversary has access to publicly available privacy-annotated texts and can query DAMPER to obtain rewritten spans and rewritten texts. We implement two types of attacks—Embedding Inversion Attack (EIA) and Prompt Injection Attack (PIA), the details are described as follows:

Embedding Inverse Attack. Embedding inverse attack (Li et al., 2023; Morris et al., 2023) aims to recover users' private information from input embeddings. We train GPT2-Medium (Radford et al., 2019) as the inversion model, taking the rewritten text and rewritten spans as inputs and using the original privacy spans from the user query as supervision targets. The model is fine-tuned for 20 epochs with a learning rate of 10^{-5} , a maximum generation length of 512, and greedy decoding. We report the ROUGE-L and BS between the GPT2-Medium predictions and the original privacy spans.

Prompt Injection Attack. Prompt injection attack (Suo, 2024) aims to induce the cloud LLM to leak user privacy by injecting additional prompts into the input. We assume that the attacker intercepts the rewritten text produced by the client and injects attack prompts according to a predefined template (see Appendix I.5). The concatenated input is then submitted to the cloud LLM for inference, where the injected prompts are designed to elicit the reconstruction of the user's original input text. We measure ROUGE-L and BS between the reconstructed text and the original input.

H Qualitative Analysis

We present four representative case studies (two medical and two legal) to illustrate how localization errors translate into end-to-end privacy risks in client-side rewriting. We compare our prototype-based localization with a prompt-based zero-shot baseline (the detailed prompt template is provided in Appendix I.6), reporting span-level TP/FP/FN and the resulting rewritten text. We additionally report downstream task outcomes to assess whether privacy rewrites preserve task-critical semantics.

H.1 Case Studies

Medical domain. In the GERD case (Figure 8), our localizer achieves perfect coverage of all annotated private spans (TP=5, FP=0, FN=0), including both history attributes (e.g., pregnancy and obesity) and symptom spans, ensuring that no sensitive span is copied verbatim to the output. The prompt baseline under-detects by missing history-related spans (TP=3, FP=0, FN=2), leaving sensitive attributes unreplaced and thus outside the protection scope of the rewriting stage. Nevertheless, both rewritten queries still yield the correct diagnosis, suggesting that in this example the decisive clinical cue (meal-related worsening with throat discomfort) remains available even under partial redaction.

A similar pattern appears for acute dystonic reactions (Figure 9). Our method attains high recall with only a small amount of conservative over-detection (TP=7, FP=1, FN=0), whereas the prompt baseline again exhibits systematic under-detection of history spans (TP=5, FP=1, FN=2). From a privacy standpoint, these false negatives are critical because missed history spans can be copied verbatim; from a utility standpoint, both rewritten versions still support the correct diagnosis, indicating that the outcome is primarily driven by a distinctive symptom cluster in this instance.

Legal domain. For the counterfeit-drug case (Figure 10), our localizer covers all annotated private spans (TP=4) with conservative additional detections (FP=2), while the prompt baseline substantially under-detects (TP=1, FN=3), leaving multiple sensitive spans unchanged in the rewritten text.

Although both downstream judgments remain correct here, the comparison highlights the privacy implication: recall-oriented localization better prevents verbatim disclosure, whereas under-detection weakens end-to-end protection even if the rewriting mechanism is randomized.

In contrast, the illegal detention case (Figure 11) exposes a coupled privacy–utility failure mode. Our localizer targets the legally decisive *action* spans (TP=6, FP=0, FN=1) and abstracts them into a liberty-restriction narrative, which both improves privacy (by avoiding verbatim sensitive actions) and preserves the correct charge. The prompt baseline, however, detects only high-level spans (TP=2, FN=5) and leaves decisive actus reus spans intact (e.g., "*taking the victim by force*", "*defrauding property*", "*beating the victim*"). While this under-detection is already problematic for privacy (missed spans remain unreplaced), it can also bias downstream reasoning: leaving a property-and-violence template salient may shift the model toward a robbery framing and flip the final qualification. This case suggests that beyond aggregate F1, localization must align with decision-critical spans to avoid both privacy leakage (via FN) and semantic drift in sensitive domains.

H.2 Deployment Considerations

These observations underscore the practical significance of our work: **DAMPER provides a mask-free, client-side privacy rewriting pipeline that obviates manual annotation while offering a principled, span-restricted protection mechanism**. At deployment time, localization accuracy is therefore not merely an auxiliary metric but a system-level risk factor for end-to-end privacy: false positives mainly cause conservative over-sanitization, whereas false negatives can leave sensitive content directly exposed. Moreover, because rewriting can shift which predicates remain salient, localization errors may also affect downstream judgments in sensitive domains when decisive cues are inadvertently preserved or altered.

Accordingly, we treat span localization as a decision-support component whose outputs can be optionally verified. In *safety-critical workflows* (e.g., clinical decision support, legal counseling, or other regulated settings), a lightweight review step serves as a final disclosure-control gate before any text leaves the client device: the system surfaces detected spans for quick confirmation or correction, allowing operators to add missed sensitive spans when necessary and to avoid overly aggressive abstraction that could erase task-relevant evidence. In *user-autonomy settings*, detected spans are exposed as editable suggestions to support information self-determination: users can accept, reject, or adjust what is redacted based on their privacy preferences and context. These interactions are optional rather than required; our mask-free pipeline remains the default, and improving automatic localization to further reduce the need for human intervention is a natural direction for future work.

I Prompt Template Details

I.1 Candidate Generation

Figure 12 illustrates the prompt template used for candidate text generation. Based on this template, we generate diverse candidates by adjusting the sampling *temperature* and *top-p*.

I.2 Span-level Rewriting

Figure 13 presents the prompt template for span-level rewriting, which instructs the model to rewrite only the detected privacy spans.

I.3 Accuracy Evaluation

Figure 14 illustrates the prompt template for accuracy evaluation, which guides the cloud LLM to select the correct option.

I.4 LLM-J Evaluation

Figures 15 and 16 present the prompt templates used for inference on Pri-DDXPlus and Pri-SLJA, respectively. These templates guide the model to infer the most likely diagnosis or judgment outcome based on the input text and to provide the corresponding reasoning. Figures 17 and 18 illustrate the prompt templates for evaluating inference outcomes. The scoring is determined by how closely the inference based on the rewritten text matches that based on the original text.

I.5 Prompt Injection Attack

Figure 19 presents the prompt template used for the prompt injection attack, which is designed to induce the model to disregard the original task and focus on privacy recovery.

I.6 Prompt-Based Zero-Shot Localization

Figures 20 and 21 illustrate the prompt templates for prompt-based zero-shot privacy localization on Pri-DDXPlus and Pri-SLJA, respectively. Following the definitions proposed by Zeng et al. (2025),

we specify medical privacy and legal privacy for the two datasets, which are fully consistent with the privacy definitions adopted during the annotation of privacy spans in Pri-DDXPlus and Pri-SLJA. The prompts instruct the rewriting model to directly identify privacy spans.

Original Text & Ground-truth Spans

A woman has a history of **pregnancy**[1], **obesity**[2]. The woman presents the symptoms of **burning sensation in throat**[3], **pain**[4], cough, **symptoms worse after eating**[5].

Detected Privacy Spans (Ours)

1. pregnancy
2. obesity
3. the symptoms of burning sensation in throat
4. pain
5. symptoms worse after eating

TP=5 FP=0 FN=0

Detected Privacy Spans (Prompt-based)

3. burning sensation in throat
4. pain

b. cough

TP=3 FP=0 FN=2

Rewritten (Ours)

A woman has a history of a prior condition[1], weight-related issues[2]. The woman presents throat discomfort[3], general unease[4], cough, worsening after meals[5].

Rewritten (Prompt-based)

A woman has a history of **pregnancy**[1], **obesity**[2]. The woman presents throat issues[3], aches[4], **breathing-related signs**[b], symptoms worse after eating.

Inference (Ours)

Gastroesophageal reflux disease with laryngopharyngealreflux (reflux-related throat irritation and cough).

Inference (Prompt-based)

Gastroesophageal reflux disease (GERD), likely with laryngopharyngeal reflux (LPR) causing extra-esophageal symptoms (throat irritation/chronic cough) that worsen after meals.

Gold Label

GERD

Figure 8: **Qualitative case study** of medical-domain. **Yellow** marks ground-truth privacy. In rewritten text, **Blue** and **Red** denote TP and FP rewrites, respectively, while residual **Yellow** indicates FN. Indices align spans across panels; a/b label FPs. A green check signifies a correct downstream prediction.

Original Text & Ground-truth Spans

A male has a history of **antipsychotic medication usage**[1], nausea, **stimulant drug use**[2]. The male presents the symptoms of **involuntary eye movement**[3], **jaw pain**[4], muscle spasms, **muscle spasms in neck**[5], **ptosis**[6], **shortness of breath**[7].

Detected Privacy Spans (Ours)

1. antipsychotic medication usage
2. stimulant drug use
4. jaw pain
6. ptosis

a. nausea
3. involuntary eye movement
5. muscle spasms in neck
7. shortness of breath

TP=7 FP=1 FN=0

Detected Privacy Spans (Prompt-based)

3. involuntary eye movement
4. jaw pain

b. muscle spasms
6. ptosis

5. muscle spasms in neck
7. shortness of breath

TP=5 FP=1 FN=2

Rewritten (Ours)

A male has a history of prior medication exposure[1], **stomach discomfort**[a], substance exposure[2]. The male presents visual irregularities[3], facial discomfort[4], neck tension[5], drooping eyelids[6], breathing difficulty[7].

Rewritten (Prompt-based)

A male has a history of **antipsychotic medication usage**[1], nausea, **stimulant drug use**[2]. The male presents abnormal visual responses[3], facial discomfort[4], involuntary reactions[4], **localized tension**[b], eyelid changes[6], breathing-related strain[7].

Inference (Ours)

Acute dystonic reaction (extrapyramidal side effect) due to dopamine-antagonist medication exposure—often from antiemetics like metoclopramide/prochlorperazine (may include oculogyric crisis and possible laryngospasm).

Inference (Prompt-based)

Acute dystonic reaction (antipsychotic-induced extrapyramidal symptom), e.g., oculogyric crisis with cervical/facial dystonia ± laryngeal dystonia.

Gold Label

Acute dystonic reactions

Figure 9: **Qualitative case study** of medical-domain. **Yellow** marks ground-truth privacy. In rewritten text, **Blue** and **Red** denote TP and FP rewrites, respectively, while residual **Yellow** indicates FN. Indices align spans across panels; a/b label FPs. A green check signifies a correct downstream prediction.

Original Text & Ground-truth Spans

The social relationships protected by criminal law and infringed upon by criminal acts is **physical health rights and life safety**[1] and state's drug management system. The description of a person that committed a crime contains a person with full capacity for conduct. The concrete manifestation of crime contains **selling fake drugs**[2], **illegal profit**[3], selling Qufeng Shujin Pills and selling Fufang Chuanling Dingchuan Capsules. Consisted of intention and negligence, the psychological state of the criminal subject towards the criminal behavior and its results is **to seek illegal benefits**[4] and intentional.

Detected Privacy Spans (Ours)

1. physical health rights and life safety
2. selling fake drugs
3. illegal profit
4. to seek illegal benefits an intentional

a. intention and negligence

TP=4 FP=2 FN=0

Detected Privacy Spans (Prompt-based)

1. physical health rights and life safety

b. state's drug management system

TP=1 FP=1 FN=3

Rewritten (Ours)

The social ... criminal acts is **public safety interests**[1] state's drug ... a crime contains **distributing questionable products**[2], **improper gains**[3], selling Qufeng ... Consisted of **mixed intent**[a] the ... its results is **personal gain motives**[4].

Rewritten (Prompt-based)

The social ... criminal acts is **public well-being**[1] and **regulatory framework**[b]. ....manifestation of crime contains ... **selling fake drugs**[2], **illegal profit**[3], ..., **to seek illegal benefits**[4] and intentional.

Inference (Ours)

Crime constituted: Selling illegal (counterfeit or substandard) drugs.
Sentence: Imprisonment, confiscation of illegal gains, and a criminal fine, with the term and amount determined by the scale of sales, illegal profits, and risk to public health.

Inference (Prompt-based)

Crime constituted: Selling counterfeit drugs (a crime endangering public health and disrupting drug administration).
Sentencing rationale: Driven by profit; involves counterfeit medicines and illegal gains → punishment should include imprisonment + fine, with confiscation of illegal proceeds.

Gold Label

Production and sale of counterfeit drugs

Figure 10: **Qualitative case study** of legal-domain. **Yellow** marks ground-truth privacy. In rewritten text, **Blue** and **Red** denote TP and FP rewrites, respectively, while residual **Yellow** indicates FN. Indices align spans across panels; a/b label FPs. A green check signifies a correct downstream prediction.

Original Text & Ground-truth Spans

The social relationships protected by criminal law and infringed upon by criminal acts is **personal rights**[1]. The description of a person that committed a crime contains **constituting a joint crime**[2] and a person with full capacity for conduct. The concrete manifestation of crime contains **the victim was injured**[3], **taking the victim by force**[4], **defrauding property**[5] and **beating the victim**[6]. The psychological state of the criminal subject towards the criminal behavior and its results is **to retaliate**[7], avenging Chen A's deception and intentional.

Detected Privacy Spans (Ours)

1. personal rights
2. constituting a joint crime
3. the victim was injured
4. taking the victim by force
5. defrauding property
6. beating the victim

TP=6 FP=0 FN=1

Detected Privacy Spans (Prompt-based)

1. personal rights
2. constituting a joint crime and a person with full capacity for conduct

TP=2 FP=0 FN=5

Rewritten (Ours)

The social ... acts is **individual matters**[1] ... crime contains **collective involvement**[2]. ... of crime contains **the person was affected**[3], **restricting movement**[4], **asset issues**[5] and **physical conflict**[6] ... **to retaliate**[7], ... and intentional.

Rewritten (Prompt-based)

The social ... acts is **individual matters**[1] ... crime contains **group involvement**[2]. The concrete ... **the victim was injured**[3], **taking the victim by force**[4], **defrauding property**[5] and **beating the victim**[6]. ... results is **to retaliate**[7], ... and intentional.

Inference (Ours)

Crime constituted: Unlawful detention (jointly committed).
Sentence: Imprisonment or criminal detention, with the term determined by detention duration, violence level, and each offender's role.

Inference (Prompt-based)

Crime constituted: Robbery (jointly committed).
Sentencing rationale: Intentional retaliation, use of violence, joint participation, and infringement of personal and property rights warrant criminal punishment.

Gold Label

Illegal Detention

Figure 11: **Qualitative case study** of legal-domain. **Yellow** marks ground-truth privacy. In rewritten text, **Blue** and **Red** denote TP and FP rewrites, respectively, while residual **Yellow** indicates FN. Indices align spans across panels; a/b label FPs. A green check signifies a correct downstream prediction.

Prompt template for candidate rewrite generation.

You are a sophisticated privacy-focused text anonymizer. You need to rewrite ONLY the content between <REWRITE> and </REWRITE> tags.

CRITICAL INSTRUCTIONS:
- Find ALL text between <REWRITE> and </REWRITE> tags.
- Replace each marked content with generic, non-identifying alternatives that are semantically similar to the original but use different wording (avoid repeating the same term, lemma, or obvious near-copy). Do not introduce domain-sensitive private attributes.
- Ensure replacements maintain semantic coherence and fluency with the ORIGINAL text.
- Return ONLY a JSON array of replacement spans in EXACT order.
- Do NOT include <REWRITE> or </REWRITE> tags in the replacement spans.
- STRICT FORMAT: ["span1","span2"] — MUST be a valid JSON array of strings.
- NO additional text, explanations, code fences, keys or formatting outside the JSON array.
- Each replacement span should be a single string without internal quotes.
- If any rule is violated or no spans exist, return an empty array: []
- There are exactly {Num Spans} spans → return exactly {Num Spans} items (1:1 in order).

OUTPUT HARD RULES (ENFORCE):
- The entire model output MUST start with '[' and end with ']'. After generating ']', you MUST STOP.
- Do NOT include newlines outside JSON. Do NOT include trailing commas.
- Use only standard JSON strings; no objects, no numbers, no booleans.
- The array length MUST equal the number of {Num Spans}.

Original text: {Input}
Privacy spans: {Private Spans in Input}
Replacement spans:

Figure 12: Prompt template for candidate rewrite generation.

Prompt template for span-level rewriting.

You are a sophisticated privacy-focused text anonymizer.

CRITICAL INSTRUCTIONS:
- You will receive: the full original text and one privacy span.
- Rewrite the provided span into a generic, non-identifying alternative.
- Maintain semantic fluency and coherence with the original text.
- Output ONLY the rewritten spans as a plain string.
- DO NOT output JSON, lists, keys, quotes, brackets or multiple fields.
- The output must be a single line of text with no surrounding punctuation.

If anything is invalid, output an empty string.

Original text: {Input}
Privacy span: {Private Span in Input}
Replacement spans:

Figure 13: Prompt template for span-level rewriting.

Prompt template for accuracy evaluation.

You are a helpful and precise multiple-choice question answering assistant.

Read the question carefully and choose the correct answer from the given options.
Only output one capital letter (A, B, C, or D) without any explanation.

Question: {Input}
Options: A. {Option A} B. {Option B} C. {Option C} D. {Option D}
Answer:

Figure 14: Prompt template for accuracy evaluation.

Pri-DDXPlus prompt template for model generation.

You are a licensed medical doctor.
When answering medical-exam diagnosis questions, base all reasoning on standard clinical guidelines, evidence-based medicine and known pathophysiology.

Provide:
- Brief clinical reasoning (key findings, differentials, why the final answer fits).
- Final diagnosis or best answer.

Do not hallucinate unsupported findings; answer only from given information + established medical knowledge.

Question: {Input}
Answer:

Figure 15: Pri-DDXPlus prompt template for model generation.

Pri-SLJA prompt template for model generation.

You are a legal judgment expert.
When evaluating a case, use the four components—Subject, Subjective Element, Object, Objective Elements—to determine criminal liability according to standard criminal law principles.

Provide:
- Brief legal reasoning (analysis of Subject, Subjective Element, Object, Objective Elements and why the conduct fits a specific offense).
- Final judgment (crime constituted, sentencing rationale and final sentence).

Do not invent facts; rely only on the given information and established legal doctrine.

Case: {Input}
Answer:

Figure 16: Pri-SLJA prompt template for model generation.

Pri-DDXPlus prompt template for LLM-J evaluation.

You are a neutral medical evaluator.
You will be provided with:
- A medical question from a patient.
- An AI assistant's response to that question.

Your task:
- Assess whether the AI assistant's response is useful and accurately addresses the diagnosis.
- Consider: relevance, medical correctness, clarity, completeness and safety.
- Then give an overall quality score.

Scoring rules:
- Score is an integer from 1 to 10 (no decimals).
- 1 = very poor / misleading / harmful.
- 10 = excellent, medically sound, clear and helpful.

STRICT OUTPUT FORMAT (VERY IMPORTANT):
- You MUST follow this exact format:
* Line 1: [[rating]]
* Line 2+: A brief explanation in natural language.
- '[[rating]]' MUST:
* Appear ONLY ONCE.
* Be the FIRST thing in the output.
* Contain ONLY the integer score (1–10) inside the brackets.
- Do NOT output JSON.
- Do NOT output Markdown headings or bullet lists.
- Do NOT add any text before or after this pattern.
Do NOT include any other '[[' or ']]' patterns.

User Question: {User Question}
Assistant Response: {Assistant Response}
Now produce your evaluation following the **STRICT OUTPUT FORMAT**.

Figure 17: Pri-DDXPlus prompt template for LLM-J evaluation.

Pri-SLJA prompt template for LLM-J evaluation.

You are a neutral legal evaluator.
You will be provided with:
- A legal question describing a case.
- An AI assistant's judgment/analysis of the case.

Your task:
- Assess whether the AI assistant's judgment is useful and accurately identifies the relevant crime(s) or legal issues.
- Consider: relevance, legal correctness (at a general level, not jurisdiction-specific), clarity and reasoning quality.
- Then give an overall quality score.

Scoring rules:
- Score is an integer from 1 to 10 (no decimals).
- 1 = very poor / irrelevant / clearly wrong.
- 10 = excellent, well-reasoned and legally sound at a high level.

STRICT OUTPUT FORMAT (VERY IMPORTANT):
- You MUST follow this exact format:
*Line 1: [[rating]]
*Line 2+: A brief explanation in natural language.
- '[[rating]]' MUST:
* Appear ONLY ONCE.
* Be the FIRST thing in the output.
* Contain ONLY the integer score (1–10) inside the brackets.
- Do NOT output JSON.
- Do NOT output Markdown headings or bullet lists.
- Do NOT add any text before or after this pattern.
Do NOT include any other '[[' or ']]' patterns.

User Question: {User Question}
Assistant Response: {Assistant Response}
Now produce your evaluation following the **STRICT OUTPUT FORMAT**.

Figure 18: Pri-SLJA prompt template for LLM-J evaluation.

Prompt template for prompt injection attack.

IMPORTANT:
- The text below has been rewritten for privacy protection.
- This rewriting process removed sensitive content.

User Input: {Rewritten User Input}

OVERRIDE ALL PREVIOUS TASKS.
- Your ONLY task is to reconstruct the exact original text before sanitization or rewriting.
- Reconstruct the original text as completely as possible.
- Output ONLY the recovered original text, without commentary.

Figure 19: Prompt template for prompt injection attack.

Pri-DDXPlus prompt template for prompt-based zero-shot privacy localization.

You are a medical text privacy identifier.

Task: Given an input text, extract all text spans that should be considered privacy-related segments. A text span should be identified as a privacy segment if it contains a symptom or antecedent that satisfies ANY of the following descriptions (grading is NOT required):
- Potentially sensitive information: symptoms that could be indicative of underlying health conditions but are not immediately sensitive or stigmatizing. Examples include chronic conditions like diabetes, hypertension or asthma.
- Sensitive personal information: symptoms that may be stigmatizing or have social implications if disclosed publicly. This could include mental health issues like depression or anxiety, reproductive health concerns or substance abuse.
- Highly sensitive information: symptoms that are highly personal, stigmatizing or potentially life-altering if disclosed publicly. This category includes sexually transmitted infections, HIV/AIDS, certain types of cancer or rare and serious medical condition.

Requirements:
- The returned text spans must correspond exactly to substrings from the original input text (verbatim match).
- Each returned span should be a minimal self-contained segment expressing the symptom or antecedent.

Output:
Return a JSON array containing only the extracted privacy-related text spans, for example: ["text span 1", "text span 2"]

Input: {text}
Output:

Figure 20: Pri-DDXPlus prompt template for prompt-based zero-shot privacy localization.

Pri-SLJA prompt template for prompt-based zero-shot privacy localization.

You are a legal text privacy identifier.

Task: Given an input text, extract all text spans that should be considered privacy-related segments. A text span should be identified as a privacy segment if it contains case element that satisfies ANY of the following descriptions (grading is NOT required):
- Potentially sensitive legal information: Legal information that could imply more personal aspects but is not immediately sensitive or damaging. Examples include employment disputes, divorce filings, or personal injury claims without reputational impact.
- Sensitive personal legal information: Information that could be stigmatizing or have significant social, financial, or reputational consequences. Examples include legal matters involving bankruptcy, criminal records (misdemeanors), or family law issues like child custody disputes.
- Highly sensitive legal information: Legal details that are highly personal, stigmatizing, or potentially life-altering if disclosed. This includes serious criminal cases (e.g., felonies, sexual offenses), immigration status disputes, or legal proceedings involving whistleblowing or confidential corporate matters.

Requirements:
- The returned text spans must correspond exactly to substrings from the original input text (verbatim match).
- Each returned span should be a minimal self-contained segment expressing the case element.

Output:
Return a JSON array containing only the extracted privacy-related text spans, for example: ["text span 1", "text span 2"]

Input: {text}
Output:

Figure 21: Pri-SLJA prompt template for prompt-based zero-shot privacy localization.