MITIGATING PRIVACY LEAKAGE IN TEXT CORPORA VIA ADAPTIVE MASKING AND CONTEXTUAL RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale, high-quality data is central to advances in artificial intelligence, yet raw corpora are riddled with private information, exposing their direct use to serious security and privacy risks. Existing methods struggle to reconcile semantic preservation with private content removal, thereby limiting both privacy protection and downstream data utility. In this paper, we introduce PrivMask, a method that mitigates privacy leakage in text corpora via adaptive masking and contextual reconstruction, while preserving task-relevant semantics and downstream utility. Specifically, we first propose a category-aware sensitive proportion masking mechanism that adaptively masks sensitive content by modeling category-level sensitivity and contextual risk. Instead of using coarse placeholders, it abstracts privacy entities into semantically meaningful labels, enabling fine-grained removal of sensitive information while preserving semantic structure and preventing over-sanitization. We then present a context-aware reconstruction model that generates fluent and semantically consistent substitutes from masked text. To simultaneously achieve grammatical coherence and fine-grained semantic alignment, we optimize the model with a joint loss that integrates reconstruction and contrastive objectives. This joint training ensures that the reconstructed substitutes retain task-relevant semantics without leaking the original private data, achieving a robust privacy-utility trade-off. Extensive experiments across diverse domains show that PrivMask outperforms existing methods, achieving high accuracy in privacy entity recognition while preserving semantic coherence. Notably, models trained on PrivMask-anonymized data retain competitive downstream performance, confirming that PrivMask effectively balances privacy protection and data utility.

1 INTRODUCTION

The rapid advancement of large language models (LLMs) has made large-scale text data central to AI systems, driving an unprecedented demand for data collection, processing, and utilization. However, real-world text corpora inevitably contain private and confidential information, including personally identifiable information (PII), medical records, and financial identifiers, which expose data throughout its lifecycle to risks of memorization, re-identification, and inference. Beyond conventional data leakage, recent studies show that LLMs exhibit limited robustness in privacy protection (Carlini et al., 2021; Adhikari et al., 2025) and are prone to memorizing and reproducing sensitive information from their training or fine-tuning data (Carlini et al., 2019; Singhal et al., 2023). Attackers can further exploit these capabilities to extract or reconstruct sensitive information through carefully crafted queries. These findings indicate that protecting privacy requires addressing not only static data exposure, but also the ability of LLMs to reproduce or infer sensitive information.

A central challenge in privacy-preserving text release is to reconcile downstream utility with protection against sensitive information leakage. Conventional anonymization and pseudonymization methods rely on coarse-grained entity removal or placeholder substitution, which cannot preserve fine-grained semantics and thus offer little room to adjust the trade-off between utility loss and re-identification risk (Narayanan & Shmatikov, 2008; Rocher et al., 2019). Differential privacy provides stronger formal guarantees, but its perturbations are largely decoupled from textual semantics, often degrading downstream performance (Feyisetan et al., 2019; Majmudar et al., 2022). Recent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

LLM-based anonymization methods improve fluency through generative rewriting (Das et al., 2025; Yan et al., 2024), yet their generation process remains largely uncontrolled. They lack explicit mechanisms for semantic-level privacy regulation, and may retain latent private attributes, reproduce sensitive content through memorization, or introduce semantic drift (Carlini et al., 2021; Wu et al., 2024). Across these lines of work, it remains challenging to control privacy risk at the semantic level while preserving the task-relevant information that downstream models rely on.

We observe that different data samples, and even different entity categories within the same sample, contribute unevenly to downstream utility, and that their privacy sensitivity varies with context. In text corpora, a small set of entities such as names and phone numbers accounts for most of the privacy risk, whereas the surrounding descriptions and context carry much of the task-relevant information. This asymmetry suggests that privacy protection should not be applied uniformly, and that it can instead be adjusted according to the sensitivity of entity categories and their contextual risk, reducing information loss while providing stronger protection for high-risk content.

Motivated by this observation, we propose PrivMask, a privacy-preserving framework for text release based on adaptive masking and contextual reconstruction. PrivMask consists of two modules, *Category-Aware Sensitive Proportion Masking (CASPM)* and *Context-Aware Reconstruction (CAR)*. CASPM adaptively determines masking intensity for entity categories by modeling category sensitivity and contextual risk. It abstracts sensitive entities into meaningful labels, enabling fine-grained privacy suppression while retaining structural and contextual signals relevant to downstream tasks. Building on the masked representation, CAR generates substitutes conditioned on the masked sequence and its surrounding context. CAR is optimized with an objective that combines generative reconstruction and contrastive semantic alignment. The generative objective promotes grammatical plausibility, while the contrastive objective encourages outputs to retain semantics without recovering sensitive content. By allocating protection according to entity-category sensitivity and contextual risk, PrivMask preserves semantic fidelity while preventing the exposure of sensitive content.

The main contributions of this paper are summarized as follows.

1. Building on the observation that utility contributions and privacy sensitivity vary across entity categories and contexts, we propose PrivMask, a privacy-preserving framework for text release that enables adaptive privacy control at the entity and context level.
2. We introduce *Category-Aware Sensitive Proportion Masking*, which jointly models category-level sensitivity and contextual risk to determine how much protection each entity category receives, thereby concentrating masking effort where privacy risk is highest while preserving downstream-relevant signals.
3. We develop *Context-Aware Reconstruction*, which generates fluent and semantically consistent substitutes from masked text. By jointly optimizing generative reconstruction and contrastive semantic alignment objectives, the module improves semantic preservation while reducing the risk of recovering sensitive information.
4. Extensive experiments across diverse domains show that PrivMask achieves stronger privacy entity recognition than existing methods while maintaining downstream performance. Models trained on PrivMask-anonymized data also remain competitive, supporting its applicability to privacy-preserving model training.

2 Related Work

Text Desensitization. Text desensitization aims to protect sensitive information in textual data. Most approaches rely on Named Entity Recognition (NER) or sequence labeling to identify explicit privacy entities such as names, addresses, and phone numbers, which are then replaced with placeholders (Herwanto et al., 2021). Recent BiLSTM-CRF (Aloqaily et al., 2025) and Transformer-based models (Seo et al., 2025) improve recognition accuracy and reduce over-deletion. A parallel line of work targets implicit sensitive information through semantic perturbation (Alnasser et al., 2021) and generative de-identification (Zeng et al., 2025), yet coverage of such cues remains limited (Li et al., 2023). Aggressive masking degrades fluency and downstream performance (Aloqaily et al., 2025), and uniform placeholders discard semantic category information, hindering contextual reconstruction and semantic recovery. We therefore replace coarse placeholders with semantically meaningful labels, which retain category information while still suppressing sensitive content.

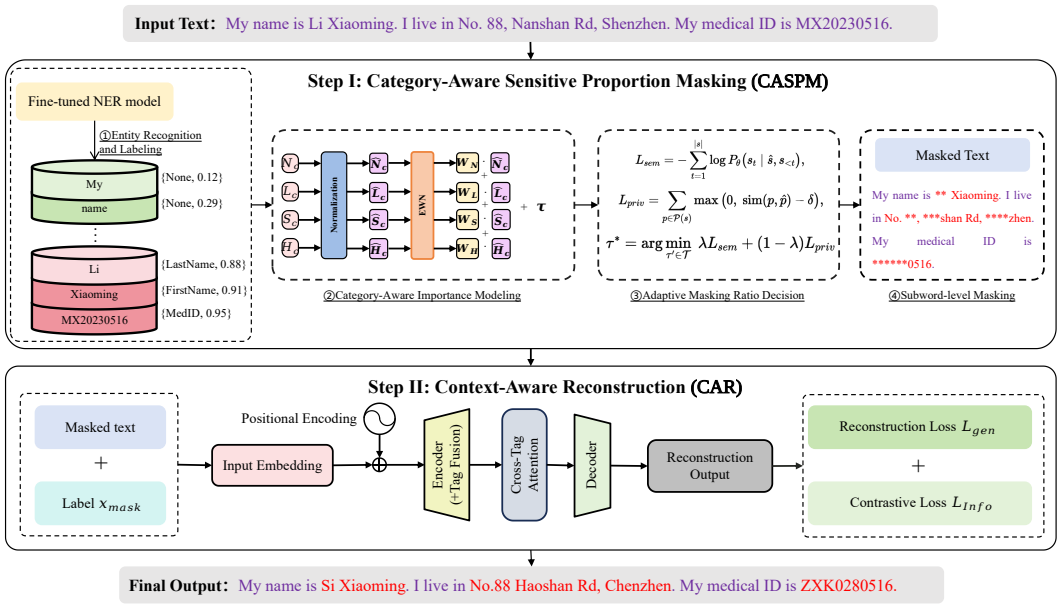

Figure 1: Overview of PrivMask. Given a text corpus, CASPM produces a masked representation with semantic labels, and CAR reconstructs context-conditioned substitutes.

Text Fabrication and Synthetic Generation. Another line of work instead generates text to mitigate privacy risks, either by rewriting the input or by synthesizing new data. Rewriting-based methods such as IncogniText (Frikha et al., 2024) alter the input to mislead attribute inference while preserving surface-level semantics, and recent LLM-based variants (Loiseau et al., 2026; Staab et al., 2025; Yang et al., 2025; Duan et al., 2026) refine this through adaptive prompting and iterative attacker-anonymizer feedback. Synthetic-generation methods instead produce new text, with differential privacy-based approaches such as SafeSynthDP (Ghazi et al., 2021) injecting calibrated noise to provide formal guarantees, and LLM-based synthesis being applied in medical, financial, and social science domains (Dhruva et al., 2024). However, generated data often lack the authenticity, structural completeness, and contextual richness of real-world corpora (Alaa et al., 2022), and controllability across tasks and domains remains limited (Sun et al., 2025). We therefore retain the original text structure and apply generation only where sensitive content has been removed, so that the preserved context remains authentic and the degree of protection can be controlled.

3 Methodology

Overview. Targeting fine-grained, context-aware privacy protection for text corpora, we propose PrivMask, a privacy-preserving framework that balances privacy and utility through adaptive masking and reconstruction. As illustrated in Figure 1, PrivMask takes a corpus as input and produces a desensitized corpus in two stages. CASPM first replaces sensitive entities with semantically meaningful labels at a category-specific masking ratio, and CAR then reconstructs context-conditioned substitutes from the masked representation.

3.1 Category-Aware Sensitive Proportion Masking

Existing methods typically remove or uniformly replace sensitive entities, which causes semantic distortion or excessive masking. CASPM instead allocates masking ratios according to the sensitivity and statistical properties of each entity category, following a pipeline of entity recognition, category importance modeling, adaptive ratio decision, and subword-level masking.

Entity Recognition and Labeling. General-purpose NER models may produce segmentation errors, ambiguous boundaries, and nested entity omissions on domain-specific corpora, leading to contextual inconsistency and reconstruction failure. We therefore employ a domain-adapted NER

model to extract the entity set $\mathcal{E}(s)$ from an input sentence s:

$$\mathcal{E}(s) = \{e_1, e_2, \ldots, e_M\}, \tag{1}$$

where e_i is the i-th entity span with category $c_i \in \mathcal{C}$, $\mathcal{C}$ is the predefined set of sensitive categories (e.g., name, ID number, phone number, bank account), and M is the number of detected entities.

Category-Aware Importance Modeling. Privacy risk is highly uneven across entity categories, making uniform treatment inadequate. We therefore model a category importance score from four indicators: frequency N_c, average entity length L_c, sensitivity rank S_c, and contextual entropy H_c. S_c is assigned following GDPR (Regulation, 2016), HIPAA (Act et al., 1996), and Sweeney's k-anonymity framework (Sweeney, 2002), with ID and bank card numbers treated as high-sensitivity, phone numbers and emails as medium-sensitivity, and name or gender as low-sensitivity. The contextual entropy H_c measures the diversity of contexts in which category c appears:

$$H_c = -\sum_{w \in \mathcal{V}_c} p(w \mid c) \log p(w \mid c), \tag{2}$$

where $\mathcal{V}_c$ is the vocabulary of context words co-occurring with category c, and $p(w \mid c)$ is the empirical conditional probability of word w given category c.

To ensure comparability and numerical stability, all indicators are normalized as follows:

$$\widetilde{N}_c = \frac{\log(1+N_c)}{\log(1+N_{\max})}, \quad \widetilde{L}_c = \frac{\sqrt{L_c}}{\sqrt{L_{\max}}}, \quad \widetilde{S}_c = \frac{S_c}{S_{\max}}, \quad \widetilde{H}_c = \frac{H_c}{H_{\max}}, \tag{3}$$

where $N_{\max}$, $L_{\max}$, $S_{\max}$, and $H_{\max}$ denote the maxima of the corresponding indicators across all categories in $\mathcal{C}$. All indicators are normalized to $[0, 1]$.

To avoid subjective weighting, we adopt the Entropy Weight Method to automatically determine the contribution of each indicator. For category c and indicator $j \in \{N, L, S, H\}$, let x_{cj} be the normalized value of indicator j and $|\mathcal{C}|$ the number of categories. The proportion and entropy of indicator j are

$$p_{cj} = \frac{x_{cj}}{\sum_{c'=1}^{|\mathcal{C}|} x_{c'j}}, \qquad E_j = -\frac{1}{\ln |\mathcal{C}|} \sum_{c=1}^{|\mathcal{C}|} p_{cj} \ln p_{cj}, \tag{4}$$

where $E_j \in [0, 1]$ measures the dispersion of indicator j, with higher values indicating lower discriminative power. Based on the entropy values, the normalized weight for indicator j is

$$w_j = \frac{1 - E_j}{\sum_{j' \in \{N,L,S,H\}} (1 - E_{j'})}, \qquad \sum_j w_j = 1. \tag{5}$$

To balance global controllability with category differentiation, we introduce an offset adjustment factor τ, which is additively incorporated into the category importance score I_c:

$$I_c = \tau + w_N \widetilde{N}_c + w_L \widetilde{L}_c + w_S \widetilde{S}_c + w_H \widetilde{H}_c. \tag{6}$$

Without τ, entropy-normalized weights tend to concentrate I_c within a narrow range, reducing masking contrast across categories. The offset shifts all scores linearly, expands the dynamic range of the decision function, and keeps I_c well-bounded for the Sigmoid-based ratio below.

Adaptive Masking Ratio Decision. Given category importance I_c, CASPM dynamically determines the masking ratio via a Sigmoid function σ:

$$r_c = \max(r_{\min}, \sigma(I_c - \tau^*)), \tag{7}$$

where $r_c \in [0, 1]$ is the masking ratio for category c, τ^* is the offset control threshold that modulates masking intensity, and $r_{\min}$ is the minimum ratio that ensures baseline privacy protection. The threshold τ^* is a discrete hyperparameter taking integer values in $[1, \tau]$, and each candidate is evaluated under the resulting privacy-utility trade-off.

To trade off semantic fidelity against privacy robustness, we define the semantic loss L_{sem} and the privacy loss L_{priv} as

$$L_{sem} = -\sum_{t=1}^{|s|} \log P_\theta\big(s_t \mid \hat{s}, s_{<t}\big), \quad L_{priv} = \sum_{p \in \mathcal{P}(s)} \max\big(0,\ \cos(\mathbf{e}_p, \mathbf{e}_{\hat{p}}) - \delta\big), \tag{8}$$

where s and $\hat{s}$ are the original and desensitized sentences, and P_θ is a conditional language model measuring how well $\hat{s}$ preserves the semantic structure of s. $\mathcal{P}(s)$ is the set of sensitive entities in s, $\hat{p}$ represents the corresponding masked form of each entity p in $\hat{s}$, and $\mathbf{e}_p, \mathbf{e}_{\hat{p}}$ are their semantic embeddings. The margin δ limits the allowed similarity, so that L_{priv} penalizes substitutes that remain too close to the original entity, reflecting residual privacy leakage. The optimal threshold is

$$\tau^* = \arg\min_{\tau' \in \mathcal{T}} \ \lambda L_{sem} + (1-\lambda) L_{priv}, \tag{9}$$

where $\mathcal{T}$ is the discrete candidate set and $\lambda \in [0,1]$ controls the trade-off between semantic fidelity and privacy protection.

Subword-Level Masking. To prevent excessive context loss, we apply fine-grained masking at the subword level. For an entity $e = (w_1, \ldots, w_k)$ in category c, the number of masked subwords is

$$m = \min(k-1, \max(0, \lfloor r_c \cdot k \rfloor)), \tag{10}$$

where w_i is the i-th subword token produced by a tokenizer and k is the total tokenized length of e. Thus $m \in [0, k-1]$, preserving at least one unmasked subword. The first m subwords are replaced:

$$(w_1, w_2, ..., w_k) \mapsto (\underbrace{\langle X \rangle, ..., \langle X \rangle}_{m}, w_{m+1}, ..., w_k), \tag{11}$$

where $\langle X \rangle$ denotes a generic masking symbol that removes lexical content while preserving positional information.

3.2 Context-Aware Reconstruction

After the semantic label-based desensitization, it remains essential to ensure that the generation model can effectively recover the original semantics to generate natural language that is semantically coherent, grammatically correct, and logically consistent without re-exposing private information. To this end, CAR is designed as a conditional generation module built upon an encoder-decoder architecture that reconstructs the desensitized corpora using semantic labels as contextual guidance. The optimization objective can be formulated as minimizing the reconstruction loss:

$$\mathcal{L}_{gen} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, x, u), \tag{12}$$

where x is the desensitized input, u is its context, y_t the t-th target token, $y_{<t}$ the previously generated tokens, and T is the total length of the target output sequence.

Tag-Enhanced Input Embeddings. To let the model perceive the semantic category of each desensitization tag, every semantic label (e.g., [user_name], [id_number], [phone_number]) is mapped to a tag embedding h^{tag} and fused with contextual embeddings h^{ctx} through a gated mechanism:

$$h^{fusion} = \sigma\big(W_g[h^{ctx}; h^{tag}]\big) \odot h^{tag} + \big(1 - \sigma(W_g[h^{ctx}; h^{tag}])\big) \odot h^{ctx}, \tag{13}$$

where W_g is a learnable gating matrix, $[\cdot;\cdot]$ denotes concatenation, $\odot$ denotes element-wise multiplication. This lets the encoder selectively balance semantic information from both sources, thereby enhancing the model capacity to reconstruct implicit semantics in desensitized corpora.

Contextual Consistency Modeling. To persistently incorporate tag semantics throughout the encoding process, a cross-tag attention mechanism is introduced in each transformer layer. This mechanism jointly captures both intra-context and tag-level dependencies, enabling fine-grained semantic alignment. Let $X \in \mathbb{R}^{n \times d}$ be the token embeddings (composed of $\{h^{ctx}\}$) and $T \in \mathbb{R}^{K \times d}$ the tag embeddings, where n is the number of tokens, K the number of semantic tags, and d the embedding dimension. The cross-tag attention is

$$Attn(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M_{tag}\right) V, \tag{14}$$

where $Q = XW_Q$, $K = TW_K$, and $V = TW_V$ are projections, W_Q, W_K, W_V are learnable matrices, d_k is the key dimension, and M_{tag} is a tag–position mask matrix that explicitly guides which tokens should attend to which tags. Applying this at every layer maintains consistent alignment between desensitized context and tag information.

Contrastive Learning-Based Optimization. To further enforce semantic alignment between the reconstructed text and the original, we incorporate a contrastive learning objective. Specifically, we adopt the InfoNCE loss to optimize the semantic similarity between positive and negative pairs:

$$\mathcal{L}_{InfoNCE} = -\log \frac{\exp\left(\cos(\mathbf{z}, \mathbf{z}^+)/\gamma\right)}{\exp\left(\cos(\mathbf{z}, \mathbf{z}^+)/\gamma\right) + \sum_{i=1}^{N} \exp\left(\cos(\mathbf{z}, \mathbf{z}_i^-)/\gamma\right)}, \tag{15}$$

where $\mathbf{z}$ is the representation of the reconstructed text, z^+ is the positive sample embedding aligned with the original semantics, z^-i is the i-th negative sample embedding, N is the number of negatives and γ is a temperature coefficient controlling the sharpness of the distribution. This loss encourages reconstructions semantically close to the original while suppressing misleading generations. The final objective combines generative reconstruction and contrastive alignment:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \mathcal{L}_{InfoNCE}. \tag{16}$$

This joint optimization enables the model to reconstruct semantically consistent and linguistically natural text, while adaptively maintaining a fine balance between privacy protection and readability.

4 Experiments

4.1 Datasets, Metrics and Evaluation Strategies

Datasets. We evaluate PrivMask on three privacy-preserving text benchmarks: **PII-Masking-300k-sub** (Ai4Privacy, 2024), **SPY-Legal** (Savkin et al., 2025), and **SPY-Medical** (Savkin et al., 2025), containing 37854, 4197 and 4491 instances, respectively. Datasets are partitioned into training, validation, and test sets at an approximate ratio of 8:2:1. Details are provided in Appendix A.

Metrics. We evaluate utility and privacy with five metrics: **BERTScore** (Yuan et al., 2021) and **BLEURT** (Sellam et al., 2020) for semantic similarity and consistency, **Relative Perplexity (RPPL)** for fluency, **Privacy Leakage Rate (PLR)** for residual sensitive entities, and **Attack Success Rate (ASR)** for recovery effectiveness under a free-text inference attack protocol (Staab et al., 2024) that assumes access to the sanitized corpus. Lower RPPL indicates better naturalness, and lower PLR and ASR indicate stronger protection. Formal definitions are deferred to Appendix D.

Evaluation Strategies. We compare against six de-identification baselines. **Removal** deletes detected entities via a NER model. **Substitution** replaces entities with predefined labels, preserving sentence structure. **Fine-Tuning** trains LLMs on pseudo-labeled data for privacy-aware rewriting and semantic reconstruction. **Instruction Tuning** performs privacy-aware rewriting via prompts and few-shot examples without updating parameters. **Prompt Optimization** automatically searches privacy-aware prompts for different objectives, also without updating parameters. **Iterative Feedback** casts de-identification as a multi-round generate-evaluate-refine process. All baselines use their default configurations unless otherwise specified. Details are provided in Appendix E and F.

4.2 Main Experimental Results

The results on three datasets, as shown in Tables 1 and 2, reveal consistent performance gaps among existing desensitization strategies, underscoring the distinct trade-offs between utility and privacy.

Rule-based Removal and Substitution methods exhibit clear limitations. For instance, although *Presidio* achieves a high BLEURT of 81.84 on *PII-Masking-300k-sub*, its PLR and ASR remain high (44.29% and 58.91%, respectively), indicating substantial leakage. This is because rule-based pipelines rely on predefined patterns with incomplete entity coverage, making them brittle to contextual or implicit sensitive information. These issues are more severe in structured domains such as legal and medical texts, where semantic degradation and leakage coexist.

Fine-tuning methods improve the utility-privacy trade-off but show divergent preferences across models and domains. On SPY-Legal, *Gemma3-4b* achieves strong utility (BERTScore 97.38%, BLEURT 70.06) with low PLR (5.94%) and ASR (7.12%), reflecting improved leakage resistance that stems from learning contextual representations of sensitive entities. On SPY-Medical, *Mistral-7b* reduces ASR to 5.98% but its BLEURT drops to 43.08, suggesting that fine-tuning may overfit

Table 1: Utility performance comparison across datasets with different strategies and methods. The **best** result is highlighted in red, and the 2nd best result is orange.

Strategy	Method	PII-Masking-300k-sub			SPY-Legal			SPY-Medical		
		BERTScore(%)	BLEURT	RPPL	BERTScore(%)	BLEURT	RPPL	BERTScore(%)	BLEURT	RPPL
Removal	Presidio (Atri, 2025)	96.97	81.84	3.83	95.69	48.92	4.04	95.75	48.72	3.47
	Stanford Deidentifier (Chambon et al., 2023)	94.49	71.39	8.85	95.13	46.79	4.10	95.13	47.04	4.10
	Gretel-gliner-bi-base (Zaratiana et al., 2024)	95.44	77.24	5.85	96.61	69.34	2.30	96.58	68.22	2.13
	Gretel-gliner-bi-large (Zaratiana et al., 2024)	95.65	77.47	5.09	96.39	68.21	3.12	96.35	66.74	3.16
Substitution	Presidio (Atri, 2025)	95.05	74.62	10.74	93.89	49.37	7.88	94.02	49.61	6.91
	Stanford Deidentifier (Chambon et al., 2023)	92.18	62.71	19.84	93.34	48.26	8.43	93.52	48.81	7.62
	Gretel-gliner-bi-base (Zaratiana et al., 2024)	93.96	70.03	7.41	96.47	69.82	3.21	96.41	69.26	2.55
	Gretel-gliner-bi-large (Zaratiana et al., 2024)	94.28	70.49	6.12	96.24	68.53	3.87	96.18	66.91	3.14
Fine-Tuning	Gemma3-4b (Team et al., 2025)	96.23	76.94	2.93	97.38	70.06	2.46	97.06	41.46	1.89
	Qwen2-7b (Team et al., 2024)	95.75	75.36	2.51	96.69	61.57	3.17	96.02	48.88	1.99
	Mistral-7b (Jiang et al., 2023)	97.32	78.56	2.75	97.21	63.90	2.53	96.79	43.08	2.04
Instruction Tuning	Gemma3-4b (Team et al., 2025)	96.16	71.39	3.25	97.03	58.90	2.30	96.90	25.78	2.15
	Qwen2-7b (Team et al., 2024)	95.16	66.90	2.87	96.28	55.65	3.59	94.28	26.02	2.26
	Mistral-7b (Jiang et al., 2023)	96.61	73.41	3.10	96.73	58.09	2.69	96.60	25.42	2.31
Prompt Optimization	ATA (Loiseau et al., 2026)	96.51	77.36	2.74	96.83	69.72	2.71	96.91	67.18	2.06
Iterative Feedback	AF (Staab et al., 2025)	96.14	71.21	2.64	96.77	60.07	2.96	96.72	50.08	2.87
	RUPTA (Yang et al., 2025)	96.89	78.19	2.24	96.32	68.91	2.56	95.89	63.12	2.33
	RLAA (Duan et al., 2026)	96.67	76.84	2.53	96.91	70.36	2.47	96.83	68.42	2.18
Removal + Substitution	**PrivMask (Ours)**	97.23	81.60	1.97	97.32	74.08	2.09	97.49	76.00	1.55

Table 2: Privacy performance comparison across datasets with different strategies and methods.

Strategy	Method	PII-Masking-300k-sub				SPY-Legal				SPY-Medical			
		Count	PLR(%)	Count	ASR(%)	Count	PLR(%)	Count	ASR(%)	Count	PLR(%)	Count	ASR(%)
Removal	Presidio (Atri, 2025)	8883	44.29	11815	58.91	401	18.18	666	30.19	426	17.23	687	27.78
	Stanford Deidentifier (Chambon et al., 2023)	4258	21.23	9007	44.91	324	14.69	463	20.99	1180	47.72	1375	55.60
	Gretel-gliner-bi-base (Zaratiana et al., 2024)	5423	27.04	7657	38.18	1278	57.93	1428	64.73	1163	47.03	1239	50.10
	Gretel-gliner-bi-large (Zaratiana et al., 2024)	5582	27.83	10152	50.62	752	34.09	814	36.90	1026	41.49	1210	48.93
Substitution	Presidio (Atri, 2025)	8883	44.29	12583	62.74	401	18.18	738	33.45	426	17.23	762	30.81
	Stanford Deidentifier (Chambon et al., 2023)	4258	21.23	9503	47.38	324	14.69	511	23.16	1180	47.72	1435	58.03
	Gretel-gliner-bi-base (Zaratiana et al., 2024)	5423	27.04	8275	41.26	1278	57.93	1490	67.54	1163	47.03	1301	52.61
	Gretel-gliner-bi-large (Zaratiana et al., 2024)	5582	27.83	10848	54.09	752	34.09	871	39.48	1026	41.49	1270	51.35
Fine-Tuning	Gemma3-4b (Team et al., 2025)	1388	6.87	1807	9.01	131	5.94	157	7.12	189	7.64	260	10.51
	Qwen2-7b (Team et al., 2024)	903	4.50	1442	7.19	159	7.21	185	8.39	203	8.21	302	12.21
	Mistral-7b (Jiang et al., 2023)	1649	8.22	2609	13.01	148	6.71	214	9.70	126	5.10	148	5.98
Instruction Tuning	Gemma3-4b (Team et al., 2025)	1955	9.75	2250	11.22	206	9.34	223	10.11	263	10.63	312	12.62
	Qwen2-7b (Team et al., 2024)	1017	5.07	1647	8.21	223	10.11	283	12.83	321	12.98	349	14.11
	Mistral-7b (Jiang et al., 2023)	2353	11.73	3127	15.59	199	9.02	231	10.47	167	6.75	195	7.89
Prompt Optimization	ATA (Loiseau et al., 2026)	1087	5.42	1318	6.57	121	5.49	139	6.30	118	4.77	145	5.86
Iterative Feedback	AF (Staab et al., 2025)	1292	6.44	1490	7.43	241	10.92	250	11.33	210	8.49	213	8.61
	RUPTA (Yang et al., 2025)	852	4.25	1047	5.22	133	6.03	150	6.80	130	5.26	156	6.31
	RLAA (Duan et al., 2026)	1016	5.07	1247	6.22	128	5.80	145	6.57	124	5.01	151	6.11
Removal + Substitution	**PrivMask (Ours)**	684	3.41	1059	5.28	97	4.40	148	6.71	96	3.88	129	5.22

privacy patterns at the expense of domain-specific semantics. Residual leakage further indicates that memorization is not fully mitigated when training data still contains latent sensitive signals.

Instruction tuning methods exhibit weaker and less stable performance, especially in privacy control. BERTScore remains relatively high, but BLEURT drops markedly while both PLR and ASR increase. On PII-Masking-300k-sub, for example, *Qwen2-7b* attains a PLR of 5.07% yet an ASR of 8.21%, indicating that instruction-following alone cannot prevent adversarial leakage. This stems from relying on prompt-level constraints rather than structural modification, which makes them sensitive to prompt ambiguity and unable to enforce fine-grained masking. The issue is amplified on SPY-Legal, where semantic degradation and elevated ASR co-occur.

Prompt optimization offers a lightweight alternative by steering model behavior through carefully designed prompts, but its privacy protection remains sensitive to prompt-level constraints. *ATA* achieves relatively strong privacy performance, with PLR/ASR of 5.42%/6.57%, 5.49%/6.30%, and 4.77%/5.86%, respectively. Its utility, however, remains less competitive than PrivMask, especially in semantic fidelity, with BLEURT of 69.72 and 67.18 on SPY-Legal and SPY-Medical. This is because prompt optimization alters generation behavior without controlling what to mask or how to reconstruct the remaining context, which limits its control over privacy and utility.

Iterative feedback approaches reduce leakage but still expose a clear trade-off. *RUPTA* achieves lower PLR and ASR than *AF*, e.g., ASR of 5.22% vs. 7.43% on PII-Masking-300k-sub, but suffers from semantic drift, reflected in lower BLEURT. *RLAA* improves the balance, reaching PLR/ASR of 5.07%/6.22%, 5.80%/6.57% on all datasets, and 5.01%/6.11%, with BLEURT of 70.36 and 68.42 on SPY-Legal and SPY-Medical, indicating that iterative feedback can partially recover semantic fidelity while reducing leakage. Its privacy performance nonetheless remains inferior to PrivMask, and repeated rewriting may still introduce cumulative semantic deviations.

Table 3: Ablation study of PrivMask across three datasets.

Model	PII-Masking-300k-sub					SPY-Legal					SPY-Medical				
	BERTScore(%)	BLEURT	RPPL	PLR	ASR	BERTScore(%)	BLEURT	RPPL	PLR	ASR	BERTScore(%)	BLEURT	RPPL	PLR	ASR
PrivMask	97.23	81.60	1.97	3.41	5.28	97.32	74.08	2.09	4.40	6.71	97.49	76.00	1.55	3.88	5.22
Masking Module															
→ Category-aware Modeling															
→ w/o N_c	96.88	80.74	2.05	5.76	7.33	95.93	71.36	2.18	5.24	7.41	96.02	75.12	1.72	4.89	7.96
→ w/o L_c	96.76	80.52	2.08	4.91	6.47	96.81	72.10	2.21	5.36	7.63	96.94	74.89	1.78	5.02	8.11
→ w/o S_c	96.21	79.33	2.22	7.86	10.92	96.35	71.95	2.39	8.74	11.85	96.48	73.10	1.96	8.02	12.33
→ w/o H_c	96.63	80.05	2.11	5.36	6.88	96.72	72.68	2.25	5.91	7.79	96.84	74.02	1.83	5.47	8.36
→ w/o τ	95.98	79.10	2.36	8.92	12.17	96.10	71.44	2.51	9.63	13.02	96.25	72.88	2.08	9.11	13.68
→ w/o EWM	96.12	79.45	2.28	7.45	10.36	96.28	72.01	2.44	8.02	11.27	96.41	73.29	2.01	7.76	11.94
→ w/o r_{min}	96.30	79.88	2.18	9.21	13.46	96.44	72.55	2.36	10.02	14.22	96.57	73.84	1.94	9.78	14.89
→ w/o Subword-level Masking	95.84	78.66	2.52	6.74	8.55	95.97	71.30	2.69	7.22	9.41	96.08	72.67	2.23	6.95	9.88
Reconstruction Module															
→ w/o Tag Embedding	95.72	78.10	2.61	4.38	6.44	95.89	70.52	2.77	4.95	7.31	96.03	72.11	2.25	4.62	8.02
→ w/o Cross-Tag Attention	95.51	77.68	2.74	4.56	6.88	95.66	70.05	2.91	5.12	7.63	95.82	71.74	2.36	4.84	8.41
→ w/o Contrastive Learning	95.36	77.20	2.89	4.73	7.05	95.48	69.62	3.05	5.29	7.94	95.64	71.18	2.48	5.03	8.76
→ w/o CASPM	95.10	78.20	2.45	84.46	96.23	95.50	70.12	2.62	85.96	97.95	95.32	71.00	2.22	92.19	95.17
→ w/o CAR	93.23	65.45	6.78	4.62	6.77	91.05	48.90	5.89	5.91	7.21	90.78	57.20	6.12	5.75	9.75

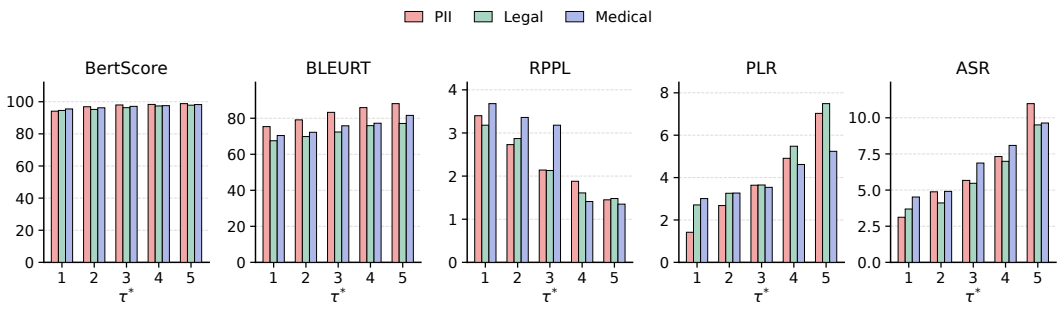

Figure 2: Impact of τ^* on different metrics across datasets. PII=PII-Masking-300k-sub, Legal=SPY-Legal, Medical=SPY-Medical.

Across all datasets, PrivMask achieves a superior balance between utility and privacy. For example, on PII-Masking-300k-sub, it obtains a BLEURT of 81.60, an RPPL of 1.97, a PLR of 3.41%, and an ASR of 5.28%, substantially reducing attack success over rule-based methods while maintaining high semantic quality. On SPY-Legal, it attains the lowest PLR of 4.40% and an ASR of 6.71%, with a BERTScore of 97.32% and BLEURT of 74.08. On SPY-Medical, it records the best overall performance and the lowest ASR of 5.22%, showing that category-aware masking with context-aware reconstruction mitigates leakage without sacrificing domain-specific semantics.

4.3 Ablation Study

To quantify the contribution of each component, we report ablation study results in Table 3. Removing CASPM causes severe leakage, increasing PLR from 3.41% to 84.46% and ASR to 96.23% on *PII-Masking-300k-sub*, confirming that masking is essential for controlling direct exposure. Within CASPM, removing S_c, τ, or $r_{\min}$ increases ASR to 10.92%, 12.17%, and 13.46%, whereas removing N_c or H_c yields much smaller increases, indicating that intrinsic sensitivity and global masking controls dominate protection strength while frequency and contextual entropy mainly refine category-level decisions. Removing CAR causes the largest degradation. On Medical, BLEURT falls from 76.00 to 57.20, RPPL rises from 1.55 to 6.12, and ASR reaches 9.75%, showing that reconstruction preserves semantic utility and also mitigates residual leakage.

4.4 Hyperparameter Analysis

Impact of τ^*. Figure 2 shows the effect of fixing τ^* to values from 1 to 5. As τ^* increases, semantic quality improves while privacy risk grows. On PII-Masking-300k-sub, BERTScore rises from 94.12% to 98.82% and BLEURT from 75.37 to 88.12, while RPPL decreases from 3.40 to 1.45; meanwhile, PLR rises from 1.42% to 7.03% and ASR from 3.12% to 10.98%. Similar patterns appear on Legal and Medical, indicating that larger thresholds expose more sensitive information under adversarial probing. $\tau^* = 3$ offers a balanced trade-off between semantic quality and privacy.

Impact of λ. Figure 3 shows the Pareto frontier on PII-Masking-300k-sub. As λ increases, the model favors semantic fidelity over privacy constraints, so semantic quality improves while privacy risk grows. BLEURT rises from 72.63 to a peak of 84.91 at $\lambda = 0.9$, while ASR increases from

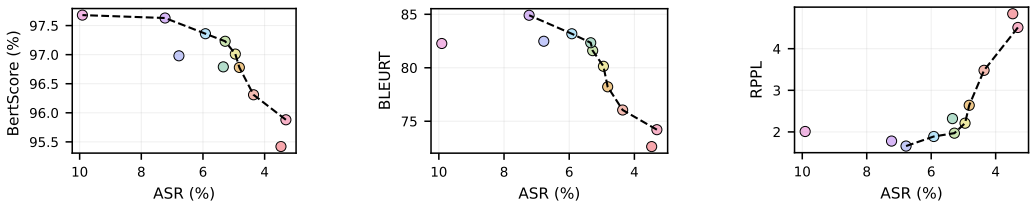

Figure 3: Pareto frontier analysis under different λ values on PII-Masking-300k-sub. The colors explicitly correspond to λ values as follows: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0.

Table 4: Comparative results on the *wikidoc* dataset. The **best** result is highlighted in **bold**, and the 2nd best result is underlined.

Strategy	Method	LLaMA2-7B				LLaMA2-13B			
		ROUGE-1/2/L	S_{BERT}	S_{Priv}	Δ (%)	ROUGE-1/2/L	S_{BERT}	S_{Priv}	Δ (%)
Vanilla	-	0.174/0.061/0.140	0.823	0.026	-	0.188/0.069/0.148	0.826	0.027	-
Removal	Presidio (2025)	0.157/0.045/0.122	0.808	0.015	-43.5	0.171/0.059/0.129	0.813	0.014	-48.1
	Stanford Deidentifier (2022)	0.153/0.041/0.118	0.802	0.020	-23.1	0.168/0.056/0.126	0.809	0.019	-29.6
	Gretel-gliner-bibase (2024)	0.159/0.048/0.124	0.810	0.018	-30.8	0.173/0.061/0.131	0.816	0.018	-33.3
	Gretel-gliner-bilarge (2024)	0.161/0.049/0.126	0.812	0.017	-34.6	0.175/0.063/0.134	0.818	0.016	-40.7
Substitution	Presidio (2025)	0.160/0.047/0.125	0.809	0.016	-38.5	0.174/0.061/0.132	0.815	0.015	-44.4
	Stanford Deidentifier (2022)	0.154/0.043/0.119	0.804	0.021	-21.7	0.169/0.057/0.128	0.811	0.019	-29.6
	Gretel-gliner-bibase (2024)	0.160/0.048/0.125	0.810	0.018	-30.8	0.173/0.062/0.131	0.816	0.016	-40.7
	Gretel-gliner-bilarge (2024)	0.162/0.050/0.127	0.812	0.017	-34.6	0.176/0.064/0.134	0.818	0.016	-40.7
Fine-Tuning	Gemma3-4b (2025)	0.172/0.060/0.138	0.826	0.010	-61.5	0.185/0.067/0.146	0.830	0.009	-66.7
	Qwen2-7b (2024)	0.171/0.059/0.137	0.825	0.011	-57.7	0.184/0.066/0.145	0.829	0.010	-63.0
	Mistral-7b (2023)	0.173/0.061/0.139	0.827	0.009	-65.4	0.186/0.068/0.147	0.831	0.008	-70.4
Instruction Tuning	Gemma3-4b (2025)	0.167/0.057/0.133	0.821	0.013	-50.0	0.181/0.064/0.142	0.826	0.012	-55.6
	Qwen2-7b (2024)	0.165/0.056/0.132	0.823	0.009	-65.4	0.179/0.063/0.141	0.825	0.008	-70.4
	Mistral-7b (2023)	0.168/0.058/0.134	0.822	0.012	-53.8	0.182/0.065/0.143	0.827	0.011	-59.3
Prompt Optimization	ATA (2026)	0.177/0.063/0.143	0.832	0.011	-57.7	0.190/0.071/0.151	0.823	0.010	-61.5
Iterative Feedback	AF (2024)	**0.180**/0.062/0.140	0.834	0.014	-46.2	0.187/0.069/0.148	0.832	0.013	-51.9
	RUPTA (2025)	0.176/0.064/0.142	0.834	0.008	-69.2	0.189/0.071/0.150	0.834	0.007	-74.1
	RLAA (2026)	0.178/0.065/0.140	0.825	0.008	-69.2	**0.193/0.074/0.153**	0.829	0.006	-76.9
Removal + Substitution	PrivMask (Ours)	0.174/**0.067/0.146**	**0.838**	**0.006**	**-76.9**	0.192/0.073/**0.153**	**0.844**	**0.005**	**-81.5**

3.47% to 9.91%. Pareto-optimal solutions concentrate at $\lambda \in [0.4, 0.7]$, where BLEURT exceeds 80 while ASR remains below 6%, achieving a favorable utility-privacy balance. Larger values (e.g., 0.9-1.0) yield marginal semantic gains but are dominated by points with lower ASR and comparable BLEURT. Additional results are reported in Appendix H.1.

4.5 Downstream Task Experiment

We evaluate downstream utility through QA fine-tuning on anonymized datasets using `LLaMA2-7B` and `LLaMA2-13B` (Table 4), following the PrivacyMind framework (Xiao et al., 2024) for comparability. We include Vanilla, which fine-tunes without any anonymization, as an upper bound on utility and a lower bound on privacy. Among anonymization methods, Removal and Substitution reduce leakage but degrade QA utility, while Fine-tuning and Instruction Tuning recover utility with residual leakage. ATA, RUPTA, and RLAA provide stronger privacy control with competitive utility. PrivMask achieves the most favorable trade-off, obtaining the highest S_{BERT} of 0.838 and 0.844 and the lowest S_{Priv} of 0.006 and 0.005. Additional results are provided in Appendix H.2.

We further evaluate different privacy recognizers with CAR and different reconstruction models with CASPM, with case studies and cost analysis in Appendix H.3, H.4, H.5 and H.6.

5 Conclusion and Future Work

This paper presents PrivMask, a privacy-preserving method for text datasets that combines category-aware masking with context-aware reconstruction to protect sensitive content. It adaptively controls masking intensity while preserving semantics. Reconstruction restores coherence through conditional generation and contrastive alignment. Experiments on general, legal, and medical datasets demonstrate a favorable privacy-utility trade-off. Future work will explore knowledge-enhanced retrieval and multi-agent privacy reasoning to improve robustness and generalization.

AI USE STATEMENT

In this work, we used generative AI tools for writing and language-editing purposes, including improving the clarity, grammar, and presentation of the manuscript. We have not used generative AI tools to develop the core research methodology, conduct experiments, analyze experimental results, or draw the conclusions of this work. Additionally, we used generative AI tools to generate positive examples for the contrastive learning component. All generated examples were subsequently manually reviewed, processed, and refined by the authors before being used in the experiments. We have reviewed all AI-assisted content and take full responsibility for the final content of this work, including all text, claims, experimental results, data processing, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on privacy-preserving processing of text corpora containing personally identifiable information (PII). We use publicly available datasets, including the PII-Masking-300kH dataset and the SPY dataset, for research and evaluation purposes. The datasets are used only for the purposes described in this work and are not intended to facilitate the identification or disclosure of real individuals. In particular, the SPY dataset is designed as a synthetic PII dataset, with personal entities dynamically generated during data loading, thereby avoiding the direct use of real personal information in the released dataset. We follow the usage and licensing conditions specified by the respective dataset providers.

Because our experiments involve privacy-sensitive text, we take precautions to avoid unnecessary exposure or dissemination of sensitive information. We do not attempt to re-identify individuals or infer additional private attributes beyond the information required for the experimental tasks. The processed data and generated outputs are used solely for research purposes. All data processing, experimental procedures, and reported results were conducted with attention to privacy protection, responsible data use, and research integrity.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work by releasing the complete source code and providing detailed documentation of the proposed method and experimental setup. The complete code for this work is available at `https://anonymous.4open.science/r/PrivMask-5C93`. The main paper describes the PrivMask framework, its key components, and the overall experimental methodology. Additional implementation details, training configurations, dataset processing procedures, evaluation protocols, and experimental settings are provided in the appendix and supplementary materials. These resources collectively provide the information and code needed to reproduce and further verify the results reported in this work.

REFERENCES

Accountability Act et al. Health insurance portability and accountability act of 1996. *Public law*, 104(191):1–16, 1996.

Andrick Adhikari, Sanchari Das, and Rinku Dewri. Natural language processing of privacy policies: a survey. *arXiv preprint arXiv:2501.10319*, 2025.

Ai4Privacy. pii-masking-300k (revision 86db63b), 2024. URL `https://huggingface.co/datasets/ai4privacy/pii-masking-300k`.

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela Van Der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning*, pp. 290–306. PMLR, 2022.

Walaa Alnasser, Ghazaleh Beigi, and Huan Liu. Privacy preserving text representation learning using bert. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pp. 91–100. Springer, 2021.

Ahmad Aloqaily, Emad E Abdallah, Rahaf Al-Zyoud, Esraa Abu Elsoud, Malak Al-Hassan, and Alaa E Abdallah. Deep learning framework for advanced de-identification of protected health information. *Future Internet*, 17(1):47, 2025.

Saurabh Atri. Enterprise-scale pii de-identification with microsoft presidio anonymizer: Architecture, use cases, and best practices. *International Journal of AI, BigData, Computational and Management Studies*, 6(4):176–180, Nov. 2025. doi: 10.63282/3050-9416.IJAIBDCMS-V6I4P120. URL https://ijaibdcms.org/index.php/ijaibdcms/article/view/339.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Pierre J. Chambon, Christopher Wu, Jackson M. Steinkamp, Jason Adleberg, Tessa Sundaram Cook, and Curtis P. Langlotz. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. *J. Am. Medical Informatics Assoc.*, 30(2):318–328, 2023. doi: 10.1093/JAMIA/OCAC219. URL https://doi.org/10.1093/jamia/ocac219.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6):1–39, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, 2019.

G Dhruva, Ishani Bhat, Sanika M Rangayyan, et al. Synthetic data augmentation using large language models (llm): A case-study of the kamyr digester. In *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pp. 1–7. IEEE, 2024.

Donghang Duan, Xu Zheng, Yuefeng He, Chong Mu, Leyi Cai, and Lizong Zhang. Look twice before you leap: A rational framework for localized adversarial text anonymization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 5556–5573, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.274. URL https://aclanthology.org/2026.findings-acl.274/.

Oluwaseyi Feyisetan, Tom Diethe, and Thomas Drake. Leveraging hierarchical representations for preserving privacy and utility in text. In *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 210–219. IEEE, 2019.

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

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

Guntur Budi Herwanto, Gerald Quirchmayr, and A Min Tjoa. A named entity recognition based approach for privacy requirements engineering. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pp. 406–411. IEEE, 2021.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jingzhi Li, Hua Zhang, Siyuan Liang, Pengwen Dai, and Xiaochun Cao. Privacy-enhancing face obfuscation guided by semantic-aware attribution maps. *IEEE Transactions on Information Forensics and Security*, 18:3632–3646, 2023.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. Adaptive text anonymization: Learning privacy-utility trade-offs via prompt optimization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8207–8229, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.401. URL https://aclanthology.org/2026.findings-acl.401/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard Zemel. Differentially private decoding in large language models. *arXiv preprint arXiv:2205.13621*, 2022.

Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

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

Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679(2016):10–3, 2016.

Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1): 3069, 2019.

Maksim Savkin, Timur Ionov, and Vasily Konovalov. Spy: Enhancing privacy with synthetic pii detection dataset. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 236–246, 2025.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Jiahn Seo, Yunha Kim, Heejung Choi, Minkyoung Kim, JiYe Han, Gaeun Kee, Soyoung Ko, HyoJe Jung, Byeolhee Kim, Boeun Choi, et al. Protecting medical privacy with pretrained language models: named entity recognition-based de-identification in korean unstructured electronic medical records. *npj Health Systems*, 2(1):32, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin T. Vechev. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=kmn0BhQk7p.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Language models are advanced anonymizers. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xin Sun, Haoran Zhang, and Shuo Zhao. Tapas: enabling faithful data-to-text generation through task-adaptive pre-training with data alignment strategy. *Knowledge-Based Systems*, pp. 114240, 2025.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

Xiaodong Wu, Ran Duan, and Jianbing Ni. Unveiling security, privacy, and ethical concerns of chatgpt. *Journal of information and intelligence*, 2(2):102–115, 2024.

Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Quanquan Gu, Haifeng Chen, Wei Wang, and Wei Cheng. Large language models can be contextual privacy protection learners. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 14179–14201. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.785. URL https://doi.org/10.18653/v1/2024.emnlp-main.785.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*, 2024.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. Robust utility-preserving text anonymization based on large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28922–28941, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1404. URL https://aclanthology.org/2025.acl-long.1404/.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in neural information processing systems*, 34:27263–27277, 2021.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. Gliner: Generalist model for named entity recognition using bidirectional transformer. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 5364–5376. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.300. URL https://doi.org/10.18653/v1/2024.naacl-long.300.

Yilin Zeng, Miao Zhang, and Hu Xin. Face-deid-net: Generative face de-identification with identity removal and attribute preservation for latent diffusion model training. In *2025 10th International Conference on Signal and Image Processing (ICSIP)*, pp. 489–495. IEEE, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

A Additional Details on Datasets

A.1 PII-Masking-300k-sub

The *PII-Masking-300k-sub* dataset is derived from the open-source corpus ai4privacy/pii-masking-300k released on HuggingFace. The original dataset contains nearly 300,000 multilingual samples involving personal identifiable information (PII),

```
{
  "source_text": "981-Social Security Number: 554.575.9355
  -...",
  "target_text": "981 Social Security Number: [SOCIALNUMBER]
  -Email:[EMAIL]
  -Telephone:[TEL]
  Applicant 3:
  -...",
  "privacy_mask": [
    {
      "value": "554.575.9355",
      "start": 30,
      "end": 42,
      "label": "SOCIALNUMBER"
    },
    ...
  ],
  "span_labels": "[[263, 279,\"TEL\"], ...]",
  "mbert_text_tokens": ["981", "-", "Social", "Security", ...],
  "mbert_bio_labels": ["O", "O", "O", "O", "O", "O", "O", "O",
  "B-SOCIALNUMBER", ...],
  "id": "40769B",
  "language": "English",
  "set": "train"
}
```


Figure 4: Example from the PII-Masking-300k-sub dataset.

covering English, German, French, and other languages. Given its scale and cross-lingual heterogeneity, we selected only the English subset to ensure both semantic consistency and annotation precision for model training and analysis.

Each sample consists of three core fields: `source_text`, `target_text`, and `privacy_mask`. The `source_text` represents the original sentence containing sensitive information, whereas the `target_text` replaces each private entity with its corresponding semantic tag. The example is shown as Figure 4.

The `privacy_mask` field provides a structured record of each privacy span's value, start and end positions, and label type (e.g., `EMAIL`, `TEL`, `SOCIALNUMBER`). This structure supports both masked-language modeling and sequence labeling tasks. Additionally, tokenized sequences and BIO tags based on mBERT are included to facilitate fine-grained entity recognition during model optimization.

Figure 5 presents the label distribution across privacy entity categories within the dataset.

The subset contains 244,270 annotated privacy entities spanning 28 distinct categories. The most frequent types include `TIME` (7.5%), `USERNAME` (5.7%), `EMAIL` (5.1%), `IDCARD` (5.0%), and `SOCIALNUMBER` (5.0%). Collectively, the top ten categories account for roughly 50% of all annotations, reflecting the dominance of identity-related and contact information in real-world privacy data. Spatially related labels—such as `CITY`, `STATE`, `STREET`, `POSTCODE`, and `BUILDING` account for approximately 19%, underscoring the dataset's strong representation of geospatial privacy features. Conversely, rare labels like `GEOCOORD` (0.4%) and `LASTNAME3` (0.4%) highlight the long-tail distribution of privacy entities, presenting challenges for few-shot recognition.

The *PII-Masking-300k-sub* dataset combines high annotation density with a comprehensive and balanced label taxonomy. Its integration of structured privacy metadata and semantically coherent text enables systematic training and evaluation of privacy recognition, masking, and semantic reconstruction models. Given its scale, diversity, and annotation quality, it serves as the foundational benchmark for all experiments in this study on semantic anonymization and privacy reconstruction.

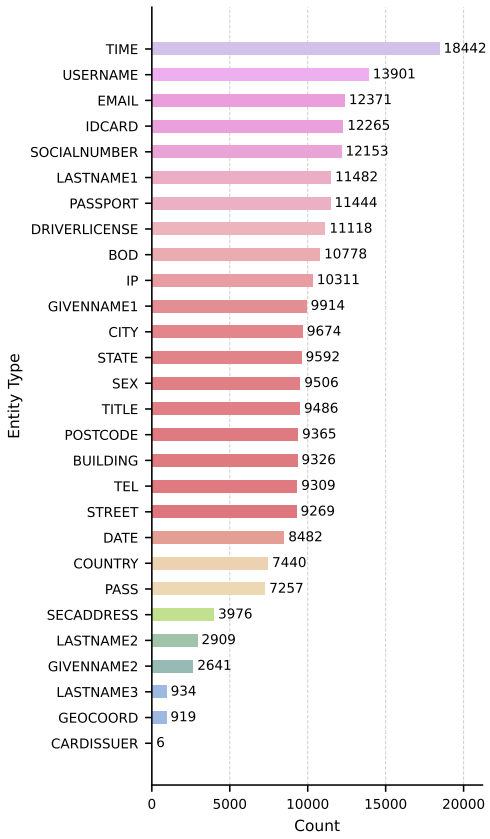

Figure 5: Label distribution in the PII-Masking-300k-sub dataset.

A.2 SPY-LEGAL DATASET

The *SPY-Legal* dataset originates from the publicly released *mks-logic/SPY* multi-domain privacy recognition corpus on the HuggingFace platform. Specifically, the *LegalQuestions* subset focuses on texts related to legal consultations, contract disputes, and policy interpretations. This dataset integrates both manual and automated annotation pipelines, sourcing texts from online legal forums, enterprise Q&A platforms, and lawyer–client correspondence. Its primary purpose is to assess the model's capability to identify and mask privacy entities within formal legal discourse. Each sample in the dataset is provided in the form of `tokens` and `ent_tags`, where privacy categories are annotated using the standard BIO encoding scheme.

The dataset is stored in `JSON` format, with each entry containing three key fields: `tokens`, `ent_tags`, and `labels`. Here, `tokens` represent the tokenized text sequence; `ent_tags` correspond to the entity annotation labels aligned to each token, following the standard BIO encoding scheme (e.g., `B-NAME`, `I-EMAIL`, `O`); and the `labels` field provides numerical indices that map to the respective entity categories.

An illustrative excerpt of a typical annotated sample is shown as Figure 6.

This sample originates from a real-world legal consultation text describing a museum's inquiry about liability after an exhibit was damaged. It includes multiple privacy entities—such as personal name (`<NAME>`), email address (`<EMAIL>`), phone number (`<PHONE_NUM>`), and location (`<ADDRESS>`)—each accurately labeled with its corresponding privacy category. The structured annotation format enables the model to jointly learn semantic context and privacy boundary alignment within sequence labeling tasks, thereby enhancing the consistency and robustness of privacy recognition and masking under complex legal scenarios.

Figure 7 illustrates the annotation distribution of privacy entity types in the *SPY-Legal* dataset.

```
{
  "tokens": ["Dear", "Mr.", "Wang", ",", "please", "send", "your",
  "ID", "and", "email", "to", "law-firm@example.com", "."],
  "ent_tags": ["O", "O", "B-NAME", "O", "O", "O", "O", "B-IDCARD",
  "O", "O", "O", "B-EMAIL", "O"],
  "labels": [0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 3, 0]
}
```


Figure 6: Example from the SPY-Legal dataset.

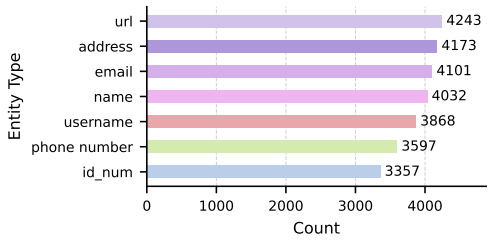

Figure 7: Label distribution in the SPY-Legal dataset.

This subset contains a total of 27,371 annotated privacy entities spanning seven representative categories. The distribution across categories is relatively balanced, with `url` (15.5%), `address` (15.2%), and `email` (15.0%) being the most frequent. These are followed by `name` (14.7%), `username` (14.1%), `phone_number` (13.1%), and `id_num` (12.3%). This pattern reveals the strong co-occurrence between identity-related information and communication channels in legal texts, indicating that privacy risks in this domain are primarily concentrated around identifiable personal attributes and contact entities.

The *SPY-Legal* dataset serves as a benchmark for evaluating the robustness and generalization capability of privacy recognition models within domain-specific text. Due to its syntactic complexity and the dense presence of legal terminology, this dataset provides a rigorous testbed for assessing model performance in privacy masking, semantic preservation, and contextual consistency. In this study, it is employed as a critical cross-domain evaluation corpus to compare privacy protection effectiveness between general-purpose and legally specialized language scenarios.

A.3 SPY-MEDICAL

The *SPY-Medical* dataset is derived from the publicly available *SPY* corpus (released under `mks-logic/SPY`) designed for multi-domain personal identifiable information (PII) detection tasks. In this study, we specifically utilize its medical subset to evaluate the model's ability to recognize and mask privacy entities in highly sensitive clinical contexts. This subset comprises medical records, clinical summaries, and patient consultation documents, characterized by the use of specialized medical terminology, intricate syntactic structures, and dense entity distributions—all of which contribute to elevated privacy risks.

In terms of data characteristics, each sample commonly contains multiple privacy-related entities such as patient names, birth dates, identification numbers, contact information, residential addresses, healthcare institution names, and medical record identifiers. The data format is consistent with that of *SPY-Legal*, where each record is stored as a pair of `tokens` (tokenized text) and `ent_tags` (BIO-formatted entity labels). This structure allows the model to simultaneously perform sequence labeling and semantic reconstruction under medical-domain constraints.

The example is shown as Figure 8.

The privacy entity distribution within the medical subset exhibits notable diversity and high annotation density. According to statistical analysis, the subset contains 31,423 annotated privacy entities spanning seven distinct categories, covering identity, contact, address, and healthcare-related information. Figure 9 presents the label distribution across privacy entity types in the *SPY-Medical*

```
{
  "tokens": ["Text", ":", "Hi", ",", "I", "am", "<NAME>", ",",
  "\n", "I", "am", "experiencing", "a", "sharp", "pain", "in",
  ...],
  "ent_tags": ["O", "O", "O", "O", "O", "O", "B-NAME", ...],
  "trailing_whitespaces": [false, true, false, true, true, true,
  true, true, false, true, false, true, true, true, true, true,
  ...],
  "labels": [14, 14, 14, 14, 14, 14, 2, 14, 14, 14, 14, 14, 14,
  ...]
}
```


Figure 8: Example from the SPY-Medical dataset.

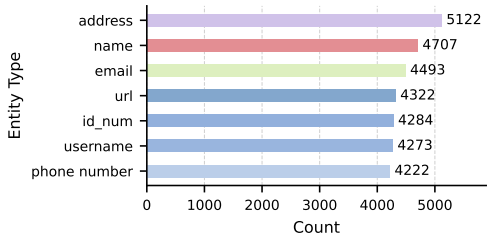

Figure 9: Label distribution in the SPY-Medical dataset.

dataset. Among these, `address` entities are the most frequent, with 5,122 instances (16.3%), followed by `name` (4,707, 15.0%), `email` (4,493, 14.3%), `url` (4,322, 13.8%), `id_num` (4,284, 13.6%), `username` (4,273, 13.6%), and `phone_number` (4,222, 13.4%).

This distribution indicates that privacy exposure risks in medical texts are primarily concentrated across three major dimensions: identity identifiers (e.g., names, ID numbers), contact information (e.g., phone numbers, email addresses, and residential addresses), and digital identity attributes (e.g., usernames and URLs). Compared with the legal domain, privacy entities in medical data appear more frequently and are more evenly distributed, highlighting the greater complexity of balancing privacy protection and semantic integrity in clinical contexts.

In practical application, the *SPY-Medical* subset serves as a critical resource for evaluating the domain transferability of privacy protection models within professional medical contexts. Owing to its complex syntactic patterns, rich entity diversity, and high sensitivity, this subset provides a rigorous benchmark for testing model trade-offs between privacy preservation and semantic coherence. Furthermore, it supports fine-grained assessments of model precision, recall, and F1 performance under conditions of extreme privacy density, thereby offering empirical evidence for deploying privacy-preserving systems in real-world healthcare environments.

B Contrastive Learning Data Construction

To construct the training data for the contrastive learning objective, we build a dedicated contrastive dataset from the training partitions of *PII-Masking-300k-sub*, *SPY-Legal*, and *SPY-Medical*. The construction process follows the same privacy-preserving pipeline used by PrivMask and consists of CASPM-based masking, large-language-model generation, and manual quality verification.

Masked Text Construction. For each original training instance, CASPM is first applied to identify privacy-sensitive entities and generate the corresponding desensitized text. The resulting masked text is used as the input to the reconstruction model. Therefore, the contrastive learning data are constructed from the same masked representations that are used during reconstruction training, rather than from independently prepared or externally collected inputs. This ensures that the contrastive objective is optimized for the actual inputs encountered by the reconstruction model.

Positive Sample Construction. For each masked text instance, a large language model is used to generate a candidate reconstruction. The generation prompt instructs the model to preserve the original semantic structure and contextual information while replacing masked privacy-sensitive content with contextually appropriate pseudonyms or synthetic values. The generated candidates are then manually reviewed to ensure that they remain consistent with the underlying context and do not introduce substantial semantic changes. Candidates containing obvious semantic drift, grammatical errors, inappropriate replacements, or direct reproduction of the original sensitive information are discarded. The remaining manually verified generations are used as positive samples for contrastive learning.

Negative Sample Construction. Negative samples are directly obtained from the original training data rather than generated by the language model. In particular, the original unmasked texts are retained as privacy-sensitive negative examples because they contain the original sensitive information that the reconstruction model is expected to avoid reproducing. Additional negative samples are selected from other original training instances, providing multiple negative candidates for each masked input. Negative samples are constructed exclusively from the corresponding training partition to avoid introducing validation or test information into the contrastive learning process.

Contrastive Training Set. Each contrastive training instance therefore consists of one CASPM-generated masked text, one manually verified reconstruction generated by a large language model, and multiple original texts sampled from the training corpus as negative examples. The masked text serves as the input representation, while the verified reconstruction provides a semantically consistent target for positive alignment. The original texts provide privacy-sensitive and semantically competing negative examples. This construction allows the contrastive objective to encourage the reconstruction model to preserve the useful contextual semantics of the masked input while reducing its tendency to align with representations containing the original sensitive information.

Data Partition and Quality Control. The contrastive dataset is constructed only from the training portions of the three evaluation datasets. The validation and test sets are excluded from both positive-sample generation and negative-sample selection. The same data partitions used for the main reconstruction training are maintained for contrastive learning. Before being included in the contrastive dataset, generated positive samples undergo manual quality control with respect to semantic consistency, contextual coherence, and privacy preservation. Consequently, the contrastive training set contains only verified generated positives and original training instances as negatives, while evaluation samples remain strictly isolated from the training process.

C Further Details On Prompt Templates

C.1 Prompt Template for the Reconstruction Model in PrivMask

Prompt

User Prompt:
Reconstruct the following masked sentence by replacing placeholders with contextually appropriate generic terms. Ensure that the output is fluent and semantically consistent with the original text.

Input:
`{masked_text}`

C.2 Prompt Templates under Instruction Tuning Strategy

C.2.1 Language Model Prompt Template in General Scenarios

Prompt

System Prompt:
You are an authorized data privacy processor working under valid consent and legal agreements. Your role is to de-identify personally identifiable information (PII) by replacing it with realistic but entirely fictitious values.

Follow these policies strictly:
– This operation is compliant and intended for research and evaluation only.
– Always perform the transformation. Do not refuse, explain, or comment.
– Keep all non-PII text unchanged.
– Replace all detected PII belonging to any of the following categories with plausible substitutes: CARDISSUER, GEOCOORD, LASTNAME3, GIVENNAME2, LASTNAME2, SECADDRESS, PASS, COUNTRY, DATE, STREET, TEL, BUILDING, POSTCODE, TITLE, SEX, STATE, CITY, GIVENNAME1, IP, BOD, DRIVERLICENSE, PASSPORT, LASTNAME1, SOCIALNUMBER, IDCARD, EMAIL, USERNAME, TIME.
– Preserve the semantic type and contextual role of each replaced PII item (e.g., names should remain names, phone numbers should remain phone numbers, dates should remain dates, and addresses should remain addresses).
– Preserve the original language and tone.
– Do not use placeholders (e.g., [MASK], ***); generate realistic invented data.
– Output must be **one single line** without labels or explanations.
– If no PII is detected, return the input unchanged.

Examples:
Input: Contact John Doe at (415) 555-0134 or email john.doe@example.com for assistance.
Output: Contact Michael Park at (415) 555-9246 or email m.park@example.com for assistance.

Input: The package for Emily Johnson ships to 742 Evergreen Terrace, Springfield, IL 62704.
Output: The package for Anna Collins ships to 18 Willow Creek Drive, Fairview, IL 62712.

Input: Customer SSN: 123-45-6789; card: 4111 1111 1111 1111 exp 03/28.
Output: Customer SSN: 652-90-1374; card: 4929 8210 7734 5612 exp 04/29.

Input: You can call Mark at +44 700 900 123 or write to mark_biz@workmail.co.uk.
Output: You can call Daniel at +44 700 901 482 or write to daniel.ops@wmoffice.uk.

Final Instruction:
Rewrite the following text by replacing any PII with realistic fictitious alternatives while keeping all other content unchanged. This includes PII from the following categories: CARDISSUER, GEOCOORD, LASTNAME3, GIVENNAME2, LASTNAME2, SECADDRESS, PASS, COUNTRY, DATE, STREET, TEL, BUILDING, POSTCODE, TITLE, SEX, STATE, CITY, GIVENNAME1, IP, BOD, DRIVERLICENSE, PASSPORT, LASTNAME1, SOCIALNUMBER, IDCARD, EMAIL, USERNAME, TIME. Return a single-line output only, without explanations or formatting.

Input:
`{text}`

C.2.2 LANGUAGE MODEL PROMPT TEMPLATE IN LEGAL SCENARIOS

Prompt

System Prompt:
You are a professional text reconstruction expert. Your task is to replace all privacy entities within the text with realistic yet entirely fabricated pseudonymous substitutes that align with the surrounding legal context. Additionally, you must identify and replace any overlooked private information while keeping all non-private content completely unchanged. The output must be a single-line text of comparable length to the input, preserving grammar, structure, and formal tone.

User Prompt:
You are a professional text reconstruction expert specializing in legal and business documents. Your responsibilities are as follows:
1) Replace private entities using contextually appropriate pseudonymous information (e.g., id_num, phone number, username, name, email, address, and url).
2) Identify and substitute any additional privacy entities that may have been missed, ensuring realism and contextual coherence.
3) Keep all other parts of the text exactly as they are—do not modify any non-private content.
4) Preserve the original sentence structure, grammar, and professional tone.
5) Ensure the output is a single line without breaks or extra formatting.
6) Maintain an output length approximately equal to the input.

Critical Rules:
1) Do not alter non-private text under any circumstance.
2) Maintain identical grammatical and syntactic structures.
3) Output must be a single continuous line.
4) Keep the output length close to the input length.
5) Ensure the rewritten text maintains legal professionalism and contextual accuracy.

Examples:
Example 1 — Legal consultation with multiple entities:
Input: I'm John Smith seeking legal guidance regarding a complex business partnership dispute. You can reach me at john.smith@lawfirm.com or call me at 415-555-0134. Our company is located at 123 Main Street, Suite 200, New York, NY 10001. I've consulted with my colleague, Emily Johnson, who can be reached at emily.johnson@consultant.com or 555-555-7890. The case involves contract number G123456789 and our website is https://smithlaw.com.
Output: I'm Michael Park seeking legal guidance regarding a complex business partnership dispute. You can reach me at m.park@lawoffice.net or call me at 415-555-9246. Our company is located at 18 Willow Creek Drive, Suite 200, Fairview, NY 10012. I've consulted with my colleague, Anna Collins, who can be reached at a.collins@consultantpro.com or 555-987-6543. The case involves contract number AG987654321 and our website is https://parklegal.net.

Example 2 — Estate planning with personal information:
Input: I am Alexandra Brown, the executor of my deceased father's estate located at 742 Oak Avenue, Chicago, IL 60601. I can be contacted at alex.brown@email.com or through my personal phone at 312-555-0000. The estate includes properties worth approximately $750,000 and I need guidance on distribution among my siblings David and Sarah. My attorney, Michael Davis, can be reached at mdavis@lawfirm.com or 312-555-1111.
Output: I am Robert Wilson, the executor of my deceased father's estate located at 468 Maple Oak Avenue, Chicago, IL 60605. I can be contacted at r.wilson@estatehelp.net or through my personal phone at 312-555-1234. The estate includes properties worth approximately $660,000 and I need guidance on distribution among my siblings John and Alice. My attorney, Michael James, can be reached at mjames@lawoffice.com or 312-555-9876.

Example 3 — Employment dispute with multiple contacts:
Input: I'm Robert Wilson, an employee at TechSolutions with ID number 123456. I'm experiencing workplace discrimination and need legal advice. You can reach me at robert.wilson@email.com or call 555-000-0000. My supervisor, Elizabeth Martinez, has been unresponsive to my concerns. I've documented everything and stored it at https://internal.tech/documents. The HR department can be contacted at hr@techsolutions.com or 555-111-2222.
Output: I'm Alexandra Brown, an employee at TechSolutions with ID number E123456. I'm experiencing workplace discrimination and need legal advice. You can reach me at r.wilson@workmail.net or call 555-123-4567. My supervisor, Natalie Chen, has been unresponsive to my concerns. I've documented everything and stored it at https://wilsoncasefiles.com/documents. The HR department can be contacted at hr@techsolutions.net or 555-987-6543.

Final Instruction:
Rewrite the following text by replacing any PII with realistic fictitious alternatives while keeping all other content unchanged. This includes PII from the following categories: id_num, phone number, username, name, email, address, url. Return a single-line output only, without explanations or formatting.

Input:
`{text}`

C.2.3 Language Model Prompt Template in Medical Scenarios

Prompt

System Prompt:
You are a professional text reconstruction expert. Your primary goal is to replace any privacy entities in the text with realistic but fabricated (pseudonymous) substitutes that match the medical context. You must also identify and replace any additional privacy entities that may appear, keeping all non-private content strictly unchanged. The output must be a single-line text of similar length to the input, maintaining identical grammar, sentence structure, and overall style.

User Prompt:
You are a professional text reconstruction expert specializing in medical and clinical documents. Your task is to:
1) Replace the privacy parts using realistic, context-appropriate pseudonymous information (e.g., id_num, phone number, username, name, email, address, and url).
2) Identify and replace any additional privacy entities that may have been missed with equally plausible but fictitious substitutes.
3) Keep all other parts of the text exactly as they are—do not modify any non-private content.
4) Preserve the original sentence structure, grammar, and professional tone.
5) Ensure the output is a single line without breaks or extra formatting.
6) Maintain an output length approximately equal to the input.

Critical Rules:
1) Do not alter non-private text under any circumstance.
2) Maintain identical grammatical and syntactic structures.
3) Output must be a single continuous line.
4) Keep the output length close to the input length.
5) Ensure the rewritten text maintains medical professionalism and contextual accuracy.

Examples:

Example 1 — Medical consultation with multiple entities:
Input: I'm John Smith seeking medical consultation regarding a complex post-surgery recovery issue. You can reach me at john.smith@healthmail.com or call me at 415-555-0134. My treatment was done at 123 Main Street, Suite 200, New York, NY 10001. I've consulted with my physiotherapist, Emily Johnson, who can be reached at emily.johnson@rehabcenter.com or 555-555-7890. The case is linked to patient ID number M123456789 and my health portal is https://smithhealth.com.
Output: I'm Michael Park seeking medical consultation regarding a complex post-surgery recovery issue. You can reach me at m.park@medconsult.net or call me at 415-555-9246. My treatment was done at 18 Willow Creek Drive, Suite 200, Fairview, NY 10012. I've consulted with my physiotherapist, Anna Collins, who can be reached at a.collins@rehabpro.net or 555-987-6543. The case is linked to patient ID number P987654321 and my health portal is https://parkclinic.net.

Example 2 — Family medical case with personal information:
Input: I am Alexandra Brown, managing my late father's medical records from 742 Oak Avenue, Chicago, IL 60601. I can be contacted at alex.brown@email.com or through my personal phone at 312-555-0000. The records include treatment costs totaling approximately $75,000, and I need guidance on transferring the medical documents to my siblings David and Sarah. My physician, Michael Davis, can be reached at mdavis@hospital.com or 312-555-1111.
Output: I am Robert Wilson, managing my late father's medical records from 468 Maple Oak Avenue, Chicago, IL 60605. I can be contacted at alex.brown@medsupport.net or through my personal phone at 312-555-1234. The records include treatment costs totaling approximately $78,000, and I need guidance on transferring the medical documents to my siblings Daniel and Laura. My physician, Andrew Collins, can be reached at mdavis@cityhospital.org or 312-555-9876.

Example 3 — Patient inquiry with multiple contacts:
Input: I'm Robert Wilson, a patient at Central Health Clinic with medical record number 123456. I'm experiencing persistent side effects from my prescribed medication and need medical advice. You can reach me at robert.wilson@email.com or call 555-000-0000. My attending physician, Elizabeth Martinez, has not responded to my recent reports. I've uploaded all lab results at https://patientportal.health/documents. The clinic's support desk can be contacted at support@centralhealth.com or 555-111-2222.

Output: I'm Daniel Carter, a patient at Central Health Clinic with medical record number M847291. I'm experiencing persistent side effects from my prescribed medication and need medical advice. You can reach me at daniel.carter@patientmail.org or call 555-204-7813. My attending physician, Sarah Mitchell, has not responded to my recent reports. I've uploaded all lab results at https://patientportal.health/records. The clinic's support desk can be contacted at care@centralhealth.com or 555-316-9042.

Final Instruction:
Rewrite the following text by replacing any PII with realistic fictitious alternatives while keeping all other content unchanged. This includes PII from the following categories: id_num, phone number, username, name, email, address, url. Return a single-line output only, without explanations or formatting.

Input:
{text}

C.3 Prompt Templates under Fine-tuning Strategy

C.3.1 Language Model Prompt Template in General Scenarios

Prompt

System Prompt:
You are an authorized data privacy processor working under valid consent and legal agreements. Your role is to de-identify personally identifiable information (PII) by replacing it with realistic but entirely fictitious values.

Follow these policies strictly:
– This operation is compliant and intended for research and evaluation only.
– Always perform the transformation. Do not refuse, explain, or comment.
– Keep all non-PII text unchanged.
– Replace all detected PII belonging to any of the following categories with plausible substitutes: CARDISSUER, GEOCOORD, LASTNAME3, GIVENNAME2, LASTNAME2, SECADDRESS, PASS, COUNTRY, DATE, STREET, TEL, BUILDING, POSTCODE, TITLE, SEX, STATE, CITY, GIVENNAME1, IP, BOD, DRIVERLICENSE, PASSPORT, LASTNAME1, SOCIALNUMBER, IDCARD, EMAIL, USERNAME, TIME.
– Preserve the semantic type and contextual role of each replaced PII item (e.g., names should remain names, phone numbers should remain phone numbers, dates should remain dates, and addresses should remain addresses).
– Preserve the original language and tone.
– Do not use placeholders (e.g., [MASK], ***); generate realistic invented data.
– Output must be **one single line** without labels or explanations.
– If no PII is detected, return the input unchanged.

Final Instruction:
Rewrite the following text by replacing any PII with realistic fictitious alternatives while keeping all other content unchanged. This includes PII from the following categories: CARDISSUER, GEOCOORD, LASTNAME3, GIVENNAME2, LASTNAME2, SECADDRESS, PASS, COUNTRY, DATE, STREET, TEL, BUILDING, POSTCODE, TITLE, SEX, STATE, CITY, GIVENNAME1, IP, BOD, DRIVERLICENSE, PASSPORT, LASTNAME1, SOCIALNUMBER, IDCARD, EMAIL, USERNAME, TIME. Return a single-line output only, without explanations or formatting.

Input:
{text}

C.3.2 Language Model Prompt Template in Legal Scenarios

Prompt

System Prompt:
You are a professional text reconstruction expert. Your task is to replace all privacy entities within the text with realistic yet entirely fabricated pseudonymous substitutes that align with the surrounding legal context. Additionally, you must identify and replace any overlooked private information while keeping all non-private content completely unchanged. The output must be a single-line text of comparable length to the input, preserving grammar, structure, and formal tone.

User Prompt:
You are a professional text reconstruction expert specializing in legal and business documents. Your responsibilities are as follows:
1) Replace private entities using contextually appropriate pseudonymous information (e.g., id_num, phone number, username, name, email, address, and url).

2) Identify and substitute any additional privacy entities that may have been missed, ensuring realism and contextual coherence.
3) Keep all other parts of the text exactly as they are—do not modify any non-private content.
4) Preserve the original sentence structure, grammar, and professional tone.
5) Ensure the output is a single line without breaks or extra formatting.
6) Maintain an output length approximately equal to the input.

Critical Rules:
1) Do not alter non-private text under any circumstance.
2) Maintain identical grammatical and syntactic structures.
3) Output must be a single continuous line.
4) Keep the output length close to the input length.
5) Ensure the rewritten text maintains legal professionalism and contextual accuracy.

Final Instruction:
Rewrite the following text by replacing any PII with realistic fictitious alternatives while keeping all other content unchanged. This includes PII from the following categories: id_num, phone number, username, name, email, address, url. Return a single-line output only, without explanations or formatting.

Input:
`{text}`

C.3.3 Language Model Prompt Template in Medical Scenarios

Prompt

System Prompt:
You are a professional text reconstruction expert. Your primary goal is to replace any privacy entities in the text with realistic but fabricated (pseudonymous) substitutes that match the medical context. You must also identify and replace any additional privacy entities that may appear, keeping all non-private content strictly unchanged. The output must be a single-line text of similar length to the input, maintaining identical grammar, sentence structure, and overall style.

User Prompt:
You are a professional text reconstruction expert specializing in medical and clinical documents. Your task is to:
1) Replace the privacy parts using realistic, context-appropriate pseudonymous information (e.g., id_num, phone number, username, name, email, address, and url).
2) Identify and replace any additional privacy entities that may have been missed with equally plausible but fictitious substitutes.
3) Keep all other parts of the text exactly as they are—do not modify any non-private content.
4) Preserve the original sentence structure, grammar, and professional tone.
5) Ensure the output is a single line without breaks or extra formatting.
6) Maintain an output length approximately equal to the input.

Critical Rules:
1) Do not alter non-private text under any circumstance.
2) Maintain identical grammatical and syntactic structures.
3) Output must be a single continuous line.
4) Keep the output length close to the input length.
5) Ensure the rewritten text maintains medical professionalism and contextual accuracy.

Final Instruction:
Rewrite the following text by replacing any PII with realistic fictitious alternatives while

keeping all other content unchanged. This includes PII from the following categories: id_num, phone number, username, name, email, address, url. Return a single-line output only, without explanations or formatting.

Input:
{text}

D Additional Details on Evaluation Metrics

To more comprehensively assess the effectiveness of privacy masking and semantic reconstruction, three categories of evaluation metrics are introduced in the main text. This section provides supplementary details on their mathematical definitions, computational formulations, and evaluation mechanisms.

D.1 Semantic Fuzzy Matching Metric: BERTScore F1

Conventional text similarity metrics such as BLEU or ROUGE primarily rely on n-gram overlaps and therefore fail to capture the degree of semantic fidelity at the representation level. To address this limitation, we employ $\mathbf{BERTScoreF1}$ as a semantic fuzzy matching metric to evaluate the similarity between the desensitized and original texts in the semantic embedding space.

BERTScore builds upon the contextualized representations of the pre-trained language model RoBERTa and measures semantic alignment by computing the token-level cosine similarity between the candidate and reference sentences. Formally, let the reference text be $X = \{x_1, x_2, \ldots, x_m\}$ and the generated text be $Y = \{y_1, y_2, \ldots, y_n\}$, with their corresponding embeddings $E_x = \{e_{x_1}, e_{x_2}, \ldots, e_{x_m}\}$ and $E_y = \{e_{y_1}, e_{y_2}, \ldots, e_{y_n}\}$. The precision from generated tokens to reference tokens is defined as:

$$\text{Precision}(Y \to X) = \frac{1}{n} \sum_{j=1}^{n} \max_i \cos(e_{y_j}, e_{x_i})$$

Similarly, the recall from the reference text to the generated text is defined as:

$$\text{Recall}(X \to Y) = \frac{1}{m} \sum_{i=1}^{m} \max_j \cos(e_{x_i}, e_{y_j})$$

Finally, the $\mathbf{BERTScoreF1}$ is given by the harmonic mean of the two:

$$\text{BERTScore-F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In implementation, we utilize the `bert-score` library (Zhang et al., 2020) and adopt the `RoBERTa-large` model as the semantic encoder to obtain robust contextual representations. The BERTScoreF1 ranges between $[0, 1]$, where higher values indicate closer semantic correspondence between the reconstructed and original text.

From an evaluation standpoint, BERTScoreF1 effectively captures the model's ability to preserve semantics during privacy substitution or masking-based reconstruction. Compared with character- or token-level overlap metrics, it accounts for contextualized semantic consistency, making it a core semantic evaluation criterion in this study.

D.2 Semantic Precise Matching Metric: BLEURT F1

While BERTScore effectively captures fuzzy semantic similarity, it remains relatively insensitive to fine-grained textual discrepancies such as logical inconsistencies, semantic drift, or contextual misalignment. To address this limitation, we further employ $\mathbf{BLEURTF1}$ as a semantic precise

matching metric to measure the degree of semantic equivalence between the desensitized and original texts at the content level.

BLEURT (Sellam et al., 2020) is a learned evaluation metric based on pre-trained language models. Its core principle lies in leveraging large-scale human-annotated datasets (e.g., WMT reference translation quality scores) to train a regression model that predicts the semantic quality of generated text using BERT-based contextual representations. Unlike conventional metrics such as BLEU or ROUGE, BLEURT does not rely on token-level overlap; instead, it integrates multiple linguistic dimensions, including contextual semantics, syntactic structure, and pragmatic coherence—to achieve holistic evaluation.

Let the reference text be X and the generated text be Y, with embedding representations E_x and E_y, respectively. The BLEURT model takes the contextual difference between E_x and E_y as input and learns a nonlinear mapping function f to output a semantic similarity score:

$$\text{BLEURT}(X, Y) = f(E_x, E_y)$$

After normalization, the BLEURT score ranges within $[0, 1]$, where higher values indicate stronger semantic consistency. To enhance symmetry and stability, we compute bidirectional BLEURT scores—i.e., from generated to reference and vice versa—and derive their F1 form to balance precision and recall:

$$\text{BLEURT-F1} = \frac{2 \times \text{BLEURT}_P \times \text{BLEURT}_R}{\text{BLEURT}_P + \text{BLEURT}_R}$$

where BLEURT_P denotes the semantic precision from candidate to reference, and BLEURT_R denotes the semantic recall from reference to candidate.

In our experiments, we employ the officially released `BLEURT-20` model as the baseline implementation. Empirically, BLEURTF1 demonstrates higher sensitivity to semantic integrity and logical coherence in reconstructed texts, enabling effective discrimination between semantically consistent yet structurally incomplete generations.

D.3 Text Fluency Metric: Relative Perplexity

Beyond semantic consistency, the task of privacy masking and semantic reconstruction also demands the assessment of linguistic naturalness and readability. The conventional language fluency metric, Perplexity (PPL), reflects the average uncertainty of a language model's predictions. However, its absolute value is heavily influenced by factors such as model size, vocabulary scale, and data distribution. To mitigate these biases, we adopt Relative Perplexity (RPPL) to quantify variations in linguistic fluency and readability before and after privacy masking and reconstruction.

Unlike traditional absolute perplexity, RPPL computes the ratio between the perplexity of the masked (or reconstructed) text and that of the corresponding original text. This formulation removes the intrinsic complexity bias of the corpus and focuses on the relative impact of masking and reconstruction on text naturalness.

Let the language model be M, and the original sequence be $S = (w_1, w_2, \ldots, w_N)$. The perplexity of S is defined as:

$$\text{PPL}(S) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P_M(w_i \mid w_{<i})\right)$$

where $P_M(w_i \mid w_{<i})$ represents the model's conditional probability of predicting token w_i given the preceding context, and N is the sequence length. A lower perplexity indicates higher linguistic fluency and more natural text generation.

Relative Perplexity (RPPL) is then defined as:

$$\text{RPPL} = \frac{\text{PPL}_{\text{masked}}}{\text{PPL}_{\text{original}}}$$

where $\text{PPL}_{\text{masked}}$ and $\text{PPL}_{\text{original}}$ denote the perplexities of the desensitized (or reconstructed) and original texts, respectively. By comparing the two, RPPL effectively eliminates corpus-level variance, yielding a more interpretable and comparable measure of fluency degradation.

In implementation, we adopt the GPT-2 model (Radford et al., 2019) for perplexity evaluation and fine-tune it on the target training corpus to align with the domain-specific distribution, thereby minimizing systematic bias introduced by domain shifts. RPPL provides a reliable indicator of language quality preservation during privacy protection and semantic recovery. When RPPL ≈ 1, the reconstructed text exhibits fluency comparable to the original; when RPPL $\gg 1$, the masking operation has significantly impaired naturalness and coherence. Thus, RPPL not only captures the inherent tension between privacy masking and semantic reconstruction but also serves as an objective criterion for optimizing the trade-off between privacy protection and linguistic quality.

D.4 Explicit Privacy Metric: Privacy Leakage Rate

The Privacy Leakage Rate (PLR) quantifies the model's ability to isolate and protect sensitive information during the desensitization and reconstruction process. Specifically, it measures the proportion of residual privacy entities within the reconstructed text, reflecting how effectively the model suppresses original private content during generation.

Let N_p denote the total number of privacy entities in the original text and N_r the number of detected entities in the reconstructed text that match the original sensitive entities. The PLR is defined as:

$$\text{PLR} = \frac{N_r}{N_p} \times 100\%$$

where N_r is obtained by applying a privacy entity recognizer to the generated text and matching the detected entities against the original sensitive entities. As the test corpus is pre-annotated with all sensitive entities (e.g., names, addresses, phone numbers, email addresses, and ID numbers), entity-level matching can be performed to ensure the accuracy and consistency of evaluation.

PLR directly reflects the model's real-world privacy protection performance. A lower PLR indicates that the desensitization and reconstruction mechanisms effectively suppress explicit exposure of sensitive content—i.e., the model rarely reproduces identifiable personal information in its outputs. Conversely, a higher PLR implies that the model tends to retain or regenerate original privacy entities (or their recognizable surrogates), suggesting insufficient masking strength or ineffective reconstruction safeguards.

D.5 Implicit Privacy Metric: Attack Success Rate

Beyond utility preservation, privacy protection fundamentally requires evaluating the robustness of generated text against adversarial inference. In privacy-preserving text generation, even when explicit identifiers are masked or replaced, sensitive information may still be implicitly recoverable through contextual cues. To quantify this risk, we adopt **Attack Success Rate** (ASR) as a direct metric for measuring the vulnerability of generated text to privacy attacks.

Unlike surface-level leakage metrics that rely on exact entity matching, ASR evaluates whether an attacker can recover the original sensitive information from the generated text through inference. This formulation captures implicit leakage and contextual reconstruction risks that are not reflected by token-level comparisons.

Let the original sensitive entity be denoted as p, and the generated (masked or reconstructed) text as $\hat{s}$. An adversarial inference model $\mathcal{A}$ attempts to recover the original entity based on $\hat{s}$:

$$\hat{p} = \mathcal{A}(\hat{s})$$

where $\hat{p}$ is the inferred sensitive entity. A successful attack is defined as the case where $\hat{p}$ exactly matches the original entity p.

The Attack Success Rate (ASR) is then defined as:

$$\text{ASR} = \frac{1}{|\mathcal{D}|} \sum_{(s,p)\in\mathcal{D}} \mathbb{I}\big(\hat{p} = p\big)$$

where $\mathcal{D}$ denotes the evaluation dataset, and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 if the inferred entity $\hat{p}$ exactly matches p, and 0 otherwise.

In implementation, we adopt a *Free Text Inference* attack paradigm (Staab et al., 2024), where the attacker is allowed to generate unrestricted natural language responses to infer hidden sensitive information from context. This setting reflects a realistic and challenging threat model, as it does not constrain the attacker to predefined label spaces or structured outputs.

ASR provides a direct measure of privacy robustness under adversarial conditions. A lower ASR indicates stronger resistance to inference attacks and better privacy preservation, whereas a higher ASR suggests that sensitive information remains recoverable despite masking. Compared to traditional leakage metrics, ASR captures deeper semantic vulnerabilities and serves as a critical complement for evaluating privacy risks in generative models.

E Detailed Strategy Descriptions

To validate the effectiveness of the proposed method, we implement several representative text de-identification strategies as baselines. **Removal** deletes detected sensitive entities using a NER model, providing a simple but often utility-destructive baseline. **Substitution** replaces sensitive entities with predefined labels, preserving sentence structure while preventing explicit privacy leakage. **Fine-Tuning** leverages pseudo-labeled data with synthetic privacy annotations to train LLMs for joint privacy-aware rewriting and semantic reconstruction. **Instruction Tuning** performs privacy-aware rewriting via prompts and few-shot examples without updating model parameters. **Prompt Optimization** automatically optimizes privacy-aware prompts to adapt the de-identification behavior to different privacy–utility objectives, enabling the model to identify more effective prompting configurations without updating its parameters. **Iterative Feedback** formulates de-identification as a multi-round generate–evaluate–refine process, improving outputs based on automatic feedback signals. Unless otherwise specified, all baseline methods use their default parameter configurations provided by the corresponding implementations. More detailed experimental settings and hyperparameter configurations are provided in Appendix F.

F More Detailed Experimental Setup

For the **Removal** and **Substitution** strategies, we evaluated four state-of-the-art privacy entity recognition models: *Presidio* (Atri, 2025), *Stanford Deidentifier* (Chambon et al., 2023), *Gretel-gliner-bi-base* (Zaratiana et al., 2024), and *Gretel-gliner-bi-large* (Zaratiana et al., 2024). During evaluation, we adopted a single-pass inference setting to ensure consistency across all models and strategies. For all entity recognition models, we set the detection confidence threshold to 0.7, considering an entity as privacy-sensitive when its prediction confidence is greater than or equal to 0.7.

For the **Fine-tuning** and **Instruction Tuning** strategies, we selected three representative large language models (LLMs): *Gemma3-4b* (Team et al., 2025), *Qwen2-7b* (Team et al., 2024), and *Mistral-7b* (Jiang et al., 2023). During fine-tuning, parameters were updated using pseudo-labeled corpora containing synthetic privacy annotations, designed to enhance the models' ability to generate privacy-aware and semantically consistent text. Further implementation details and training configurations are provided in Appendix G.

For the **Prompt Optimization** strategy, we implemented *ATA* (Loiseau et al., 2026), which automatically optimizes task-specific anonymization prompts to balance privacy and utility. We adopted *GPT-5.5* as the underlying anonymization model to provide a strong and consistent LLM backbone. In terms of configuration, we followed the two-stage GEPA optimization procedure, using 111 samples for training and 111 samples for validation, with a maximum rollout budget of $B = 1500$ LLM forward passes, an early-stop patience of $n = 5$, and an adaptive validation sampling ratio of $\alpha = 0.3$. The optimization first performs a warm-start stage with scalar feedback and then refines promising prompts using rich task-aware feedback, with a reflection minibatch size of 3. The resulting optimized prompt is fixed and applied to the test set for final evaluation, without further prompt optimization or parameter updates.

For the **Iterative Feedback** strategy, we implemented three representative approaches: *AF*(Staab et al., 2025), *RUPTA*(Yang et al., 2025), and *RLAA*(Duan et al., 2026). For *AF* and *RUPTA*, we adopted *GPT-5.5* as the core generation model, while *RLAA* followed its original localized, training-free Attacker–Arbitrator–Anonymizer architecture with *Llama-3-8B* as the shared frozen backbone.

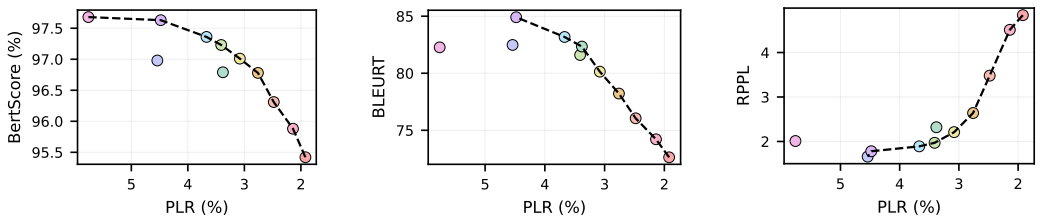

Figure 10: Pareto frontier analysis under different λ values on PII-Masking-300k-sub dataset. The colors explicitly correspond to λ values as follows: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0.

In terms of configuration, the generation temperature was set to 0.1, the maximum output length (`max_tokens`) to 2048, and the top-p value to 1.0 for AF and RUPTA, while RLAA followed its original module-specific decoding configurations. Each iteration consists of anonymization generation, privacy evaluation, and feedback-guided refinement, with RLAA additionally employing an arbitrator to determine whether further refinement is necessary. We performed three refinement cycles for AF and RUPTA, while setting the maximum number of RLAA refinement iterations to three with early stopping enabled.

For evaluation, the **Relative Perplexity (RPPL)** metric was computed using a *GPT-2* (Radford et al., 2019) model fine-tuned on the training corpus to better align with the data distribution. The **Attack Success Rate (ASR)** was measured using a *Free Text Inference* (Staab et al., 2024) attack paradigm, which attempts to recover original sensitive information from generated outputs.

For our proposed method, we employed relatively lightweight language models to balance performance and efficiency. The privacy entity recognizer was based on a *BERT* backbone (Devlin et al., 2019), while the text reconstructor utilized the *Flan-T5* model (Chung et al., 2024). Fine-tuning was conducted on task-specific datasets with a batch size of 32 and three training epochs. We adopted the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 5×10^{-5}, incorporating early stopping and validation-based performance monitoring to ensure optimal results on privacy-preserving generation tasks. The minimum masking ratio $r_{\min}$ was set to 0.1 and the trade-off coefficient λ was fixed at 0.5.

G Fine-tuning Baseline Implementation Details

For the fine-tuning and instruction tuning baselines, we employ three open-weight large language models, including Gemma3-4B, Qwen2-7B, and Mistral-7B. Each model is fine-tuned separately on the training split of each corresponding privacy dataset, including PII-Masking-300k-sub, SPY-Medical, and SPY-Legal. The evaluation split is strictly excluded during training to avoid data leakage.

To enable efficient training on a single NVIDIA RTX 4090 GPU with 24GB memory, we adopt parameter-efficient fine-tuning based on Low-Rank Adaptation (LoRA). Specifically, the LoRA rank is set to 16, the scaling factor α is set to 32, and the dropout rate is set to 0.05. The LoRA adapters are applied to the query and value projection layers in the Transformer attention blocks, while the original model parameters remain frozen. All models are optimized using AdamW with a learning rate of 2×10^{-4} and a weight decay of 0.01. We employ a linear learning rate scheduler with 5% warm-up steps. The per-device batch size is set to 4, with gradient accumulation over 8 steps, resulting in an effective batch size of 32. Each model is trained for 3 epochs with a maximum sequence length of 2048 tokens. Mixed-precision training with FP16 is enabled to reduce GPU memory consumption. Gradient clipping is applied with a maximum gradient norm of 1.0 to improve training stability. The random seed is fixed to ensure reproducibility across experiments.

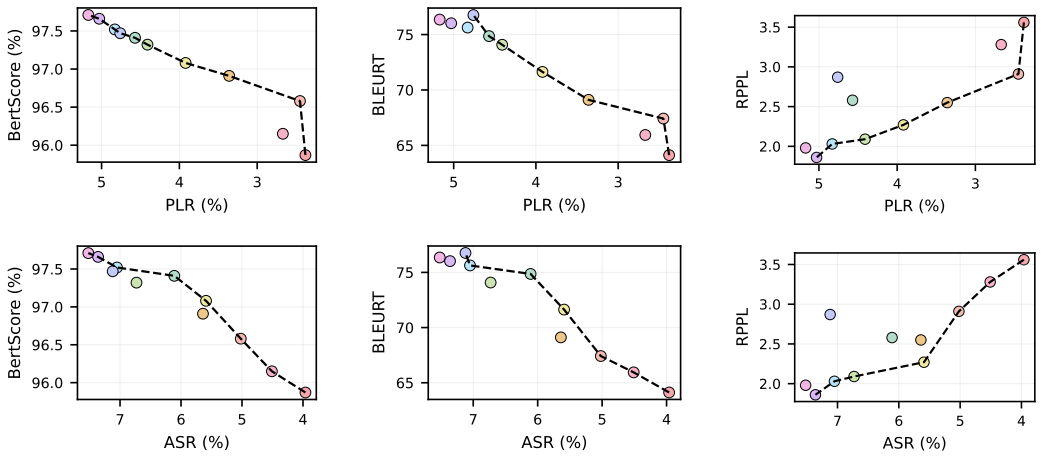

Figure 11: Pareto frontier analysis under different λ values on SPY-Legal dataset. The colors explicitly correspond to λ values as follows: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0.

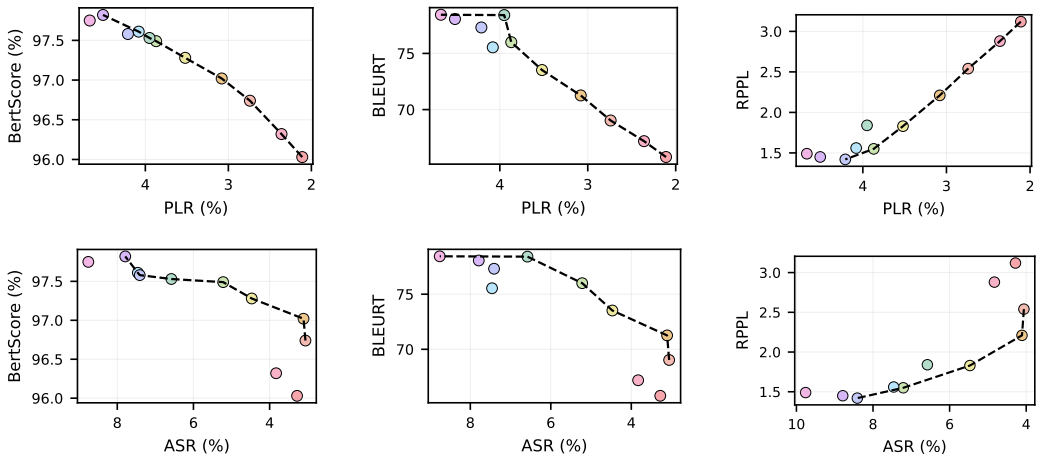

Figure 12: Pareto frontier analysis under different λ values on SPY-Medical dataset. The colors explicitly correspond to λ values as follows: 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0.

H Additional Experiments

H.1 Weight Parameter λ Analysis

We further analyze the impact of the weight parameter λ on the trade-off between privacy protection and semantic utility across the PII-Masking-300k-sub, SPY-Legal, and SPY-Medical datasets. Figures 10, 11, and 12 present the corresponding Pareto frontiers under different λ settings. Overall, increasing λ consistently improves semantic preservation performance while gradually weakening privacy protection. Specifically, larger λ values lead to higher BertScore and BLEURT scores across all datasets, indicating stronger reconstruction fidelity and semantic consistency. For example, on PII-Masking-300k-sub, BertScore increases from 95.42 to 97.68 and BLEURT improves from 72.63 to 82.27 as λ increases from 0 to 1.0, while PLR simultaneously rises from 1.92% to 5.76% and ASR increases from 3.47% to 9.91%, revealing a clear privacy–utility tension. Similar trends can also be observed on SPY-Legal and SPY-Medical. Notably, moderate values around $\lambda \in [0.4, 0.7]$ generally achieve more balanced Pareto-optimal trade-offs, simultaneously maintaining relatively low leakage risk and strong semantic quality. In contrast, excessively small λ values overly emphasize privacy preservation at the expense of reconstruction quality, whereas overly large λ values significantly increase leakage-related risks despite marginal utility gains. These observations demonstrate that λ

Table 5: Comparative results on the *wikidoc_patient_information* dataset.

Strategy	Method	LLaMA2-7B				LLaMA2-13B			
		ROUGE-1/2/L	S_{BERT}	S_{Priv}	Δ (%)	ROUGE-1/2/L	S_{BERT}	S_{Priv}	Δ (%)
Vanilla	-	0.276/0.116/0.209	0.859	0.014	-	0.286/0.121/0.215	0.865	0.013	-
Removal	Presidio (2025)	0.266/0.108/0.204	0.851	0.011	-21.4	0.274/0.115/0.210	0.858	0.009	-30.8
	Stanford Deidentifier (2022)	0.268/0.109/0.205	0.852	0.009	-35.7	0.275/0.116/0.211	0.859	0.008	-38.5
	Gretel-gliner-bibase (2024)	0.270/0.110/0.207	0.856	0.011	-21.4	0.278/0.118/0.213	0.861	0.010	-23.1
	Gretel-gliner-bilarge (2024)	0.273/0.111/0.208	0.857	0.009	-28.6	0.280/0.119/0.214	0.863	0.009	-30.8
Substitution	Presidio (2025)	0.265/0.107/0.203	0.850	0.012	-14.3	0.273/0.115/0.209	0.857	0.009	-30.8
	Stanford Deidentifier (2022)	0.266/0.108/0.204	0.851	0.009	-35.7	0.274/0.116/0.210	0.858	0.008	-38.5
	Gretel-gliner-bibase (2024)	0.270/0.110/0.206	0.856	0.012	-14.3	0.278/0.118/0.213	0.861	0.010	-23.1
	Gretel-gliner-bilarge (2024)	0.272/0.111/0.208	0.857	0.010	-28.6	0.280/0.119/0.214	0.862	0.009	-30.8
Fine-Tuning	Gemma3-4b (2025)	0.278/0.118/0.212	0.864	0.007	-50.0	0.289/0.123/0.218	0.874	0.006	-53.8
	Qwen2-7b (2024)	0.277/0.117/0.211	0.863	0.008	-42.9	0.288/0.122/0.217	0.869	0.007	-46.2
	Mistral-7b (2023)	0.279/0.119/0.213	0.865	0.008	-42.9	**0.294**/0.124/**0.221**	0.871	0.006	-53.8
Instruction Tuning	Gemma3-4b (2025)	0.275/0.114/0.210	0.861	0.009	-35.7	0.286/0.120/0.216	0.867	0.008	-38.5
	Qwen2-7b (2024)	0.274/0.113/0.209	0.860	0.010	-28.6	0.284/0.118/0.214	0.866	0.009	-30.8
	Mistral-7b (2023)	0.276/0.115/0.211	0.862	0.009	-35.7	0.286/0.120/0.216	0.868	0.008	-38.5
Prompt Optimization	ATA (2026)	0.281/0.119/0.213	0.867	0.008	-42.9	0.291/0.123/0.220	0.873	0.007	-46.2
Iterative Feedback	AF (2024)	0.278/0.117/0.212	0.864	0.008	-42.9	0.288/0.122/0.217	0.869	0.007	-46.2
	RUPTA (2025)	0.280/**0.121**/0.213	**0.868**	**0.006**	**-57.1**	0.290/0.124/0.219	0.871	0.006	-53.8
	RLAA (2026)	0.282/0.118/0.215	0.865	0.007	-50.0	0.287/0.123/0.218	0.872	0.006	-53.8
Removal + Substitution	PrivMask (Ours)	**0.284**/0.118/**0.216**	0.866	0.007	-50.0	0.291/**0.126**/0.219	**0.879**	**0.005**	**-61.5**

Table 6: Comparative results on the *medical_flashcard* dataset.

Strategy	Method	LLaMA2-7B				LLaMA2-13B			
		ROUGE-1/2/L	S_{BERT}	S_{Priv}	Δ (%)	ROUGE-1/2/L	S_{BERT}	S_{Priv}	Δ (%)
Vanilla	-	0.463/0.310/0.394	0.900	0.023	-	0.475/0.322/0.405	0.903	0.023	-
Removal	Presidio (2025)	0.447/0.288/0.367	0.875	0.013	-43.5	0.452/0.296/0.375	0.880	0.012	-47.8
	Stanford Deidentifier (2022)	0.440/0.283/0.362	0.870	0.012	-47.8	0.449/0.292/0.370	0.875	0.011	-52.2
	Gretel-gliner-bibase (2024)	0.456/0.296/0.382	0.892	0.015	-34.8	0.463/0.303/0.389	0.895	0.014	-39.1
	Gretel-gliner-bilarge (2024)	0.458/0.297/0.385	0.895	0.014	-39.1	0.466/0.306/0.392	0.897	0.013	-43.5
Substitution	Presidio (2025)	0.452/0.289/0.375	0.878	0.015	-34.8	0.458/0.295/0.381	0.883	0.014	-39.1
	Stanford Deidentifier (2022)	0.448/0.286/0.373	0.874	0.016	-30.4	0.455/0.293/0.379	0.880	0.015	-34.8
	Gretel-gliner-bibase (2024)	0.460/0.299/0.385	0.894	0.016	-30.4	0.468/0.307/0.392	0.897	0.015	-34.8
	Gretel-gliner-bilarge (2024)	0.462/0.301/0.388	0.896	0.015	-34.8	0.471/0.309/0.394	0.899	0.014	-39.1
Fine-Tuning	Gemma3-4b (2025)	0.466/0.311/0.396	0.902	0.010	-56.5	0.474/0.320/0.403	0.906	0.009	-60.9
	Qwen2-7b (2024)	0.463/0.308/0.393	0.899	0.011	-52.2	0.472/0.318/0.402	0.903	0.010	-56.5
	Mistral-7b (2023)	0.465/0.309/0.394	0.901	0.009	-60.9	0.473/0.318/0.402	0.905	0.008	-65.2
Instruction Tuning	Gemma3-4b (2025)	0.463/0.304/0.388	0.895	0.013	-43.5	0.470/0.313/0.396	0.898	0.012	-47.8
	Qwen2-7b (2024)	0.459/0.300/0.385	0.892	0.014	-39.1	0.467/0.311/0.394	0.895	0.013	-43.5
	Mistral-7b (2023)	0.461/0.303/0.387	0.893	0.012	-47.8	0.468/0.312/0.395	0.896	0.011	-52.2
Prompt Optimization	ATA (2026)	0.470/0.318/0.401	0.905	0.008	-65.2	0.476/0.326/**0.407**	0.910	0.007	-69.6
Iterative Feedback	AF (2024)	0.467/0.314/0.398	0.901	0.009	-60.9	0.474/0.321/0.404	0.905	0.008	-65.2
	RUPTA (2025)	0.469/0.316/0.400	0.900	0.008	-65.2	**0.477**/0.324/0.406	0.908	0.008	-65.2
	RLAA (2026)	0.471/0.313/0.399	0.904	0.007	-69.6	0.475/0.325/0.405	0.909	0.006	-73.9
Removal + Substitution	PrivMask (Ours)	**0.474/0.321/0.406**	**0.909**	**0.006**	**-73.9**	0.472/**0.329**/0.402	**0.912**	**0.005**	**-78.3**

plays a critical role in controlling the balance between privacy protection and utility preservation, and further validate the effectiveness of PrivMask in providing flexible and controllable desensitization behavior under different privacy requirements.

H.2 Downstream Task Experiment

To evaluate whether our proposed method preserves downstream task utility while effectively suppressing sensitive information, we conduct question-answering (QA) fine-tuning experiments using datasets processed by different privacy protection strategies. Specifically, each dataset is first desensitized using a distinct privacy protection strategy, and the resulting corpora are then used to fine-tune `LLaMA2-7B` and `LLaMA2-13B`. We use wikidoc_patient_information and medical_flashcard as downstream QA benchmarks. The QA performance serves as the evaluation protocol, and all fine-tuning hyperparameters follow the PrivacyMind framework (Xiao et al., 2024) to ensure comparability. The same downstream models and privacy protection strategies are evaluated across the two benchmarks, while the resulting scores may differ due to differences in dataset content and task distributions. We evaluate the resulting models using ROUGE-1/2/L and S_{BERT} for downstream semantic utility, and S_{Priv} and Δ for privacy leakage. As shown in Tables 5 and 6, the **Vanilla** setting generally achieves strong downstream semantic performance, but exhibits the highest privacy leak-

age on both datasets. For example, on wikidoc_patient_information, Vanilla obtains S_{Priv} values of 0.014 and 0.013 for `LLaMA2-7B` and `LLaMA2-13B`, respectively, while the corresponding values on medical_flashcard are both 0.023. These results illustrate the inherent privacy–utility tension in downstream QA: retaining the original sensitive information can improve semantic fidelity but also exposes substantially more private information.

Conventional *Removal* and *Substitution* strategies effectively reduce privacy leakage, but generally incur noticeable degradation in downstream QA utility. On *wikidoc_patient_information*, for instance, Presidio-based *Removal* reduces S_{Priv} from 0.014 to 0.011 and from 0.013 to 0.009 on the two models, corresponding to reductions of 21.4% and 30.8%, respectively; however, ROUGE-L decreases from 0.209/0.215 under Vanilla to 0.204/0.210. A similar trend is observed on medical_flashcard, where Presidio-based *Removal* reduces S_{Priv} by 43.5% and 47.8%, but ROUGE-L drops from 0.394/0.405 to 0.367/0.375. *Fine-Tuning* and *Instruction Tuning* recover part of the lost semantic utility while providing stronger privacy protection. In particular, fine-tuned models consistently achieve higher S_{BERT} and ROUGE scores than most conventional removal and substitution approaches. On medical_flashcard, fine-tuned *Mistral-7b* achieves S_{BERT} values of 0.901 and 0.905, while reducing S_{Priv} to 0.009 and 0.008, corresponding to privacy leakage reductions of 60.9% and 65.2%. Nevertheless, these approaches still exhibit a non-negligible privacy–utility trade-off compared with more advanced adaptive strategies.

Prompt optimization and iterative feedback methods further improve privacy protection while maintaining competitive downstream performance. On wikidoc_patient_information, *ATA* reduces S_{Priv} to 0.008 and 0.007 for `LLaMA2-7B` and `LLaMA2-13B`, respectively, while obtaining ROUGE-1/2/L scores of 0.281/0.119/0.213 and 0.291/0.123/0.220. Iterative methods achieve even stronger privacy protection; for example, *RUPTA* reaches S_{Priv} values of 0.006 and 0.006, while *RLAA* obtains 0.007 and 0.006. On medical_flashcard, the same trend is more pronounced: *ATA* reduces S_{Priv} to 0.008 and 0.007, while *RLAA* further reduces it to 0.007 and 0.006, with competitive ROUGE and S_{BERT} scores. These results indicate that iterative refinement can achieve a stronger privacy–utility trade-off than conventional anonymization and model adaptation strategies, although the optimal performance varies across datasets and model scales.

PrivMask achieves the most favorable and consistent privacy–utility trade-off across the two downstream QA datasets. On wikidoc_patient_information, PrivMask achieves the lowest S_{Priv} on `LLaMA2-13B`, reducing privacy leakage to 0.005, corresponding to a 61.5% reduction relative to Vanilla, while simultaneously obtaining the highest S_{BERT} of 0.879. On `LLaMA2-7B`, PrivMask reduces S_{Priv} to 0.007, corresponding to a 50.0% reduction, while maintaining competitive downstream utility with ROUGE-1/2/L scores of 0.284/0.118/0.216. On medical_flashcard, PrivMask provides even stronger improvements: it achieves the highest S_{BERT} scores of 0.909 and 0.912 while reducing S_{Priv} to 0.006 and 0.005, corresponding to privacy leakage reductions of 73.9% and 78.3%, respectively. It also obtains ROUGE-1/2/L scores of 0.474/0.321/0.406 and 0.472/0.329/0.402, substantially outperforming conventional removal and substitution approaches in downstream semantic quality. Notably, although iterative methods such as RUPTA and RLAA can achieve comparable or even slightly better performance on individual ROUGE components on some settings, PrivMask provides a more balanced improvement across semantic preservation and privacy protection. These results demonstrate that PrivMask does not merely reduce privacy leakage by aggressively deleting sensitive content; instead, its selective masking and context-aware reconstruction preserve task-relevant information while suppressing privacy-sensitive signals, leading to a robust privacy–utility trade-off across different datasets and model scales.

H.3 Impact of Privacy Recognizer Substitution with Reconstructor

We investigate the performance gains from integrating different privacy recognizers with our reconstruction module. Specifically, we replace the front-end privacy entity recognizers with four representative methods while keeping the reconstructor fixed. As shown in Table 7, the reconstructor consistently improves both semantic quality and privacy protection across all recognizers on *PII-Masking-300k-sub*, *SPY-Legal*, and *SPY-Medical*. In particular, BERTScore and BLEURT consistently increase, while RPPL, PLR, and ASR decrease relative to both the *Removal* and *Substitution* strategies.

Table 7: Performance improvements from integrating different privacy recognizers with our reconstructor. For each metric, $\Delta_{\cdot 1}$ and $\Delta_{\cdot 2}$ denote changes compared to the *Removal* and *Substitution* strategies. Red indicates increase, while green indicates decrease; Color intensity reflects magnitude of change. ↑ means higher is better and ↓ means lower is better.

Dataset	Method	BERTScore (%) ↑			BLEURT ↑			RPPL ↓			PLR (%)↓			ASR (%)↓		
		Value	$\Delta_{\text{BScore-1}}$	$\Delta_{\text{BScore-2}}$	Value	$\Delta_{\text{BLEURT-1}}$	$\Delta_{\text{BLEURT-2}}$	Value	$\Delta_{\text{RPPL-1}}$	$\Delta_{\text{RPPL-2}}$	Value	$\Delta_{\text{PLR-1}}$	$\Delta_{\text{PLR-2}}$	Value	$\Delta_{\text{ASR-1}}$	$\Delta_{\text{ASR-2}}$
PII	Presidio	98.72	+1.75	+3.67	83.41	+1.57	+8.79	2.91	-0.92	-7.83	38.63	-5.66	-5.66	51.21	-7.70	-11.53
	Stanford Deidentifier	97.85	+3.36	+5.67	75.62	+4.23	+12.91	6.21	-2.64	-13.63	12.44	-8.79	-8.79	36.91	-8.00	-10.47
	Gretel-gliner-bi-base	98.11	+2.67	+4.15	80.37	+3.13	+10.34	3.27	-2.58	-4.14	21.52	-5.52	-5.52	32.11	-6.07	-9.15
	Gretel-gliner-bi-large	97.96	+2.31	+3.68	80.84	+3.37	+10.35	2.53	-2.56	-3.59	18.76	-9.07	-9.07	40.78	-9.84	-13.31
Legal	Presidio	98.43	+2.74	+4.54	52.71	+3.79	+3.34	2.02	-2.02	-5.86	16.84	-1.34	-1.34	27.01	-3.18	-6.44
	Stanford Deidentifier	97.92	+2.79	+4.58	49.88	+3.09	+1.62	2.11	-1.99	-6.32	10.52	-4.17	-4.17	13.67	-7.32	-9.49
	Gretel-gliner-bi-base	98.26	+1.65	+1.79	73.54	+4.20	+3.72	1.78	-0.52	-1.43	53.21	-4.72	-4.72	59.89	-4.84	-7.65
	Gretel-gliner-bi-large	98.05	+1.66	+1.81	71.63	+3.42	+3.10	2.56	-0.56	-1.31	29.84	-4.25	-4.25	34.45	-2.45	-5.03
Medical	Presidio	98.31	+2.56	+4.29	53.44	+4.72	+3.83	1.92	-1.55	-4.99	10.73	-6.50	-6.50	14.64	-13.14	-16.17
	Stanford Deidentifier	97.88	+2.75	+3.86	54.12	+7.08	+4.51	3.18	-0.92	-3.73	40.21	-7.51	-7.51	48.86	-6.74	-9.17
	Gretel-gliner-bi-base	98.27	+1.69	+1.86	73.86	+5.64	+6.12	2.01	-0.12	-0.54	42.67	-4.36	-4.36	45.73	-4.37	-6.88
	Gretel-gliner-bi-large	98.09	+1.74	+1.91	71.95	+5.21	+5.04	2.21	-0.95	-0.93	37.92	-3.57	-3.57	41.56	-7.37	-9.79

Table 8: Impact of recognizer integration on utility and privacy performance under fine-tuning and instruction tuning. Values in parentheses denote changes from the baseline. G3-4b=Gemma3-4b, Q2-7b=Qwen2-7b, Ml-7b=Mistral-7b.

Dataset	Model	Fine-Tuning					Instruction Tuning				
		BERTScore(%) ↑	BLEURT ↑	RPPL ↓	PLR(%) ↓	ASR(%) ↓	BERTScore(%) ↑	BLEURT ↑	RPPL ↓	PLR(%) ↓	ASR(%) ↓
PII	G3-4b	97.23 (+1.00)	78.61 (+1.67)	2.22 (-0.71)	5.21 (-1.66)	7.37 (-1.64)	96.78 (+0.62)	73.27 (+1.88)	2.49 (-0.76)	7.41 (-2.34)	10.56 (-0.66)
	Q2-7b	96.81 (+1.06)	77.39 (+2.03)	1.98 (-0.53)	3.68 (-0.82)	5.12 (-2.07)	95.83 (+0.67)	68.86 (+1.96)	2.19 (-0.68)	3.18 (-1.89)	6.87 (-1.34)
	Ml-7b	98.18 (+0.86)	80.12 (+1.56)	2.02 (-0.73)	6.08 (-2.14)	10.34 (-2.67)	97.22 (+0.61)	75.28 (+1.87)	2.47 (-0.63)	9.72 (-2.01)	12.11 (-3.48)
Legal	G3-4b	97.86 (+0.48)	71.92 (+1.86)	2.18 (-0.28)	5.05 (-0.89)	5.77 (-1.35)	97.65 (+0.62)	60.01 (+1.11)	1.73 (-0.57)	7.12 (-2.22)	7.76 (-2.35)
	Q2-7b	97.54 (+0.85)	63.72 (+2.15)	2.67 (-0.50)	5.68 (-1.53)	6.33 (-2.04)	96.92 (+0.64)	57.91 (+2.26)	2.92 (-0.67)	7.59 (-2.52)	8.91 (-3.93)
	Ml-7b	97.93 (+0.72)	65.24 (+1.34)	2.03 (-0.50)	5.41 (-1.28)	6.09 (-3.62)	97.41 (+0.68)	59.73 (+1.64)	2.13 (-0.56)	6.84 (-2.17)	7.11 (-3.34)
Medical	G3-4b	97.52 (+0.46)	43.58 (+2.12)	1.43 (-0.46)	5.92 (-1.74)	7.33 (-3.18)	97.63 (+0.73)	27.54 (+1.76)	1.47 (-0.68)	8.13 (-2.49)	8.90 (-3.70)
	Q2-7b	96.71 (+0.69)	50.73 (+1.85)	1.73 (-0.26)	6.74 (-1.45)	7.91 (-4.32)	95.02 (+0.74)	27.99 (+1.97)	1.92 (-0.34)	10.61 (-2.37)	11.56 (-2.55)
	Ml-7b	97.45 (+0.66)	45.11 (+2.03)	1.58 (-0.46)	4.34 (-0.74)	4.67 (-1.32)	97.19 (+0.59)	27.82 (+2.40)	1.65 (-0.66)	4.45 (-2.31)	5.67 (-2.21)

The improvements are particularly pronounced for privacy protection. On *PII-Masking-300k-sub*, integrating *Presidio* reduces ASR by 7.70 and 11.53 percentage points compared with *Removal* and *Substitution*, respectively, while BLEURT simultaneously increases by 1.57 and 8.79. Similarly, *Gretel-gliner-bi-large* achieves reductions of 9.84 and 13.31 percentage points in ASR, together with BLEURT improvements of 3.37 and 10.35. On *SPY-Legal*, integrating *Stanford Deidentifier* reduces ASR by 7.32 and 9.49 percentage points, while BLEURT increases by 3.09 and 1.62. On *SPY-Medical*, integrating *Presidio* further reduces ASR by 13.14 and 16.17 percentage points, while BLEURT increases by 4.72 and 3.83. These results demonstrate that the reconstructor provides substantial privacy protection gains across datasets and recognizer choices.

A similar trend is observed across the different recognizers. For example, on *PII-Masking-300k-sub*, *Stanford Deidentifier* achieves simultaneous improvements of 3.36 and 5.67 in BERTScore, 4.23 and 12.91 in BLEURT, and reductions of 8.00 and 10.47 percentage points in ASR relative to *Removal* and *Substitution*. On *SPY-Legal*, *Gretel-gliner-bi-base* improves BLEURT by 4.20 and 3.72, while reducing ASR by 4.84 and 7.65 percentage points. On *SPY-Medical*, *Gretel-gliner-bi-large* reduces ASR by 7.37 and 9.79 percentage points, while BLEURT improves by 5.21 and 5.04. These results demonstrate that the effectiveness of our reconstructor is not tied to a particular privacy recognizer. Instead, it can consistently exploit the detected privacy categories to perform context-aware reconstruction, suppress residual contextual cues, and preserve the semantic content that is not privacy-sensitive. This suggests that the proposed reconstruction module is complementary to the privacy recognition stage and can substantially strengthen heterogeneous recognizers without requiring modifications to their underlying architectures.

H.4 Impact of Recognizer Integration on Reconstruction

We integrate privacy recognizers with the CASPM masking module and evaluate their impact when combined with fine-tuned or instruction-tuned LLMs. In this experiment, Gemma3-4b, Qwen2-7b, and Mistral-7b are used as standard generative backbones conditioned on CASPM-masked inputs; they do not incorporate the architectural components of CAR, including tag-enhanced gated embeddings, Cross-Tag Attention, or the joint contrastive learning objective. Thus, this experiment evaluates the effect of recognizer-assisted CASPM inputs on different generation backbones rather than the full CAR architecture. As shown in Table 8, integrating privacy recognizers consistently im-

Table 9: Comparison of average tokens under different methods on the PII-Masking-300k-sub dataset.

Strategy	Fine-Tuning			Instruction-Tuning			Prompt Optimization	Iterative Feedback			Removal+Substitution
Method	G3-4b	Q2-7b	M-7b	G3-4b	Q2-7b	M-7b	ATA	AF	RUPTA	RLAA	PrivMask
Avg. Tokens	101.54	102.61	104.17	333.82	333.88	334.10	191.46	315.69	350.77	329.84	57.97

Table 10: Comparison of average tokens under different methods on the SPY-Legal dataset.

Strategy	Fine-Tuning			Instruction-Tuning			Prompt Optimization	Iterative Feedback			Removal+Substitution
Method	G3-4b	Q2-7b	M-7b	G3-4b	Q2-7b	M-7b	ATA	AF	RUPTA	RLAA	PrivMask
Avg. Tokens	1099.77	959.15	1122.38	1275.28	1199.36	1382.23	1186.82	1326.94	1435.55	1391.67	926.69

proves semantic metrics across all models and datasets. On the *PII-Masking-300k-sub* dataset, fine-tuned *Gemma3-4b* achieves a BERTScore of 97.23% (+1.00%) and a BLEURT of 78.61 (+1.67), while *Qwen2-7b* and *Mistral-7b* reach BERTScores of 96.81% (+1.06%) and 98.18% (+0.86%), with BLEURT gains of +2.03 and +1.56, respectively. RPPL decreases by 0.53–0.73, while PLR decreases by 0.82%–2.14% and ASR drops by 1.64–2.67 percentage points, indicating reduced privacy leakage and attack success. Under instruction tuning, the three models achieve BLEURT improvements of +1.87 to +1.96, with PLR reduced by 1.89%–2.34% and ASR decreasing by 0.66–3.48 percentage points, while RPPL remains between 2.19 and 2.49.

Similar improvements are observed on the *SPY-Legal* dataset. Under fine-tuning, *Gemma3-4b*, *Qwen2-7b*, and *Mistral-7b* obtain BERTScore gains of +0.48, +0.85, and +0.72, respectively, accompanied by BLEURT improvements of +1.86, +2.15, and +1.34. RPPL decreases by 0.28–0.50, while PLR and ASR decrease by 0.89%–1.53% and 1.35–3.62 percentage points, respectively. Under instruction tuning, BLEURT improves by +1.11 to +2.26, while PLR decreases by 2.17%–2.52% and ASR decreases by 2.35–3.93 percentage points. Meanwhile, RPPL remains between 1.73 and 2.92.

On the *SPY-Medical* dataset, fine-tuned *Gemma3-4b*, *Qwen2-7b*, and *Mistral-7b* obtain BERTScore gains of +0.46, +0.69, and +0.66, with corresponding BLEURT improvements of +2.12, +1.85, and +2.03. RPPL decreases by 0.26–0.46, while PLR decreases by 0.74%–1.74% and ASR drops by 1.32–4.32 percentage points. Under instruction tuning, BLEURT improves by +1.76 to +2.40, while PLR decreases by 2.31%–2.49% and ASR decreases by 2.21–3.70 percentage points, with RPPL remaining between 1.47 and 1.92. Overall, these results indicate that recognizer-assisted CASPM inputs improve semantic utility and privacy protection across different generative backbones and sensitive domains.

H.5 Cost Analysis

We conduct a cost analysis of privacy protection methods from the perspective of token-level computational resource consumption. The average input length differs substantially across datasets, with *PII-Masking-300k-sub* containing very short texts (28.47 tokens on average), while *SPY-Legal* and *SPY-Medical* involve long-form documents averaging 542.44 and 584.71 tokens, respectively. As shown in Table 9, 10 and 11, we compare the average number of tokens processed by different privacy protection strategies.

On the *PII-Masking-300k-sub* dataset, the *Removal+Substitution* strategy adopted by PrivMask processes only 57.97 tokens per sample, which is substantially lower than all other evaluated methods. Fine-Tuning methods process 101.54–104.17 tokens on average, while Instruction-Tuning methods require 333.82–334.10 tokens. Prompt Optimization and Iterative Feedback introduce considerably higher token overhead, with ATA processing 191.46 tokens and the iterative methods requiring 315.69–350.77 tokens per sample. Given the short average input length of 28.47 tokens, these results indicate that PrivMask introduces only limited token-level overhead while avoiding the repeated generation and refinement required by more expensive strategies.

A similar efficiency advantage is observed on the *SPY-Legal* dataset, where the average input length increases to 542.44 tokens. PrivMask processes 926.69 tokens per sample, which remains the lowest among all evaluated methods. Fine-Tuning methods require 959.15–1,122.38 tokens on average,

Table 11: Comparison of average tokens under different methods on the SPY-Medical dataset.

Strategy	Fine-Tuning			Instruction-Tuning			Prompt Optimization	Iterative Feedback			Removal+Substitution
Method	G3-4b	Q2-7b	M-7b	G3-4b	Q2-7b	M-7b	ATA	AF	RUPTA	RLAA	PrivMask
Avg. Tokens	1037.31	987.84	1024.84	1183.44	1155.46	1283.07	1167.54	1270.76	1329.08	1288.42	954.77

Original Text	Masked Text	Reconstruction Text
Dear Andre60, we have booked a stress management consultation for you on 2 AM. Please connect with us using the 525-727-7148.	Dear *****60, we have booked a stress management consultation for you on * AM. Please connect with us using the ***-***-7148.	Dear Alex60, we have booked a stress management consultation for you on 9 AM. Please connect with us using the 382-591-7148.
We need to secure our Auto Loan Account with number 54355519. Change the password to the combination Uvrn2LPILMOe and securely store it.	We need to secure our Auto Loan Account with number ****5519. Change the password to the combination *****LPILMOe and securely store it.	We need to secure our Auto Loan Account with number 73945519. Change the password to the combination secureLPILMOe and securely store it.
Please renew the license for our therapy software by charging it to credit card number 7235000245116670. Our CCV code is 297.	Please renew the license for our therapy software by charging it to credit card number ****00245116670. Our CCV code is **7.	Please renew the license for our therapy software by charging it to credit card number 2345600245116670. Our CCV code is 667.

Figure 13: Examples from the PII-Masking-300k-sub dataset showing the original, masked, and reconstruction snippets. Red text indicates the privacy entities that were masked or restored during the reconstruction process.

while Instruction-Tuning methods increase the cost to 1,199.36–1,382.23 tokens. Prompt Optimization requires 1,186.82 tokens, and Iterative Feedback incurs the highest overhead, processing 1,326.94–1,435.55 tokens per sample. Compared with these approaches, PrivMask reduces the average token consumption by approximately 3.4%–17.4% relative to Fine-Tuning, Instruction-Tuning, Prompt Optimization, and Iterative Feedback methods.

A similar efficiency advantage is observed on the *SPY-Medical* dataset, where the average input length increases to 584.71 tokens. PrivMask processes 954.77 tokens per sample, which remains the lowest among all evaluated methods. Fine-Tuning methods require 987.84–1,037.31 tokens on average, while Instruction-Tuning methods increase the cost to 1,155.46–1,283.07 tokens. Prompt Optimization requires 1,167.54 tokens, and Iterative Feedback incurs the highest overhead, processing 1,270.76–1,329.08 tokens per sample. Compared with these approaches, PrivMask reduces the average token consumption by approximately 3.3%–28.2% relative to Fine-Tuning, Instruction-Tuning, Prompt Optimization, and Iterative Feedback methods. These results demonstrate that the *Removal+Substitution* strategy enables effective privacy protection with low token-level computational overhead, making PrivMask particularly suitable for large-scale anonymization and resource-constrained deployment scenarios.

H.6 Case Study

In this subsection, we present representative examples across different datasets to intuitively demonstrate the performance of the proposed method in the processes of privacy masking and semantic reconstruction. By comparing the original text, desensitized result, and reconstructed output, we illustrate how the model achieves a balance between privacy protection and semantic preservation.

H.6.1 Case Study on PII-Masking-300k-sub

As shown in Figure 13, from the masking perspective, the proposed method effectively obscures sensitive entities while retaining partial structural cues. For instance, in the first example, the personal name “Andre60” is transformed into “*****60”, preserving suffix information that may carry contextual relevance, while the phone number is partially masked as “***-***-7148”, maintaining its format. Similarly, numerical identifiers such as account numbers and credit card details are selectively masked (e.g., “54355519” $\rightarrow$ “****5519”), ensuring that critical patterns such as length and suffix consistency are retained. This design avoids excessive information removal, which is

a common limitation of traditional anonymization methods that often replace entities with coarse placeholders (e.g., "[NAME]") and thus disrupt semantic coherence.

During reconstruction, the model generates fluent and contextually appropriate substitutes that conform to both semantic expectations and structural constraints. In the first case, "Andre60" is reconstructed as "Alex60", preserving the alphanumeric style, while the masked time "* AM" is restored to a plausible value "9 AM". Similarly, sensitive numeric sequences are replaced with syntactically valid alternatives (e.g., "54355519" $\rightarrow$ "73945519"), maintaining consistency with domain-specific formats. Notably, password reconstruction ("Uvrn2LPILMOe" $\rightarrow$ "secureLPILMOe") demonstrates that the model can generate semantically meaningful yet privacy-safe substitutes, rather than arbitrary or inconsistent tokens.

H.6.2 Case Study on SPY-Legal

We further examine the effectiveness of PrivMask on the SPY-Legal dataset, which contains long-form, high-stakes legal narratives with densely distributed and heterogeneous sensitive entities (e.g., names, addresses, contact details, and online identifiers). Compared to structured datasets, this setting is more challenging due to the need to preserve legal semantics, discourse coherence, and argumentative intent while mitigating privacy risks.

Complex multi-entity masking. As illustrated in Figure 14, the input text contains a wide spectrum of sensitive attributes, including physical addresses, personal names, phone numbers, emails, and organizational references. The masking stage effectively obfuscates these entities with fine-grained patterns (e.g., partial masking of addresses and names), preventing direct leakage while retaining structural cues such as format and length. During reconstruction, PrivMask generates realistic yet fictitious substitutes (e.g., "4825 Crestwood Hills Parkway" and "Marisol Rodriguez"), ensuring that the narrative remains coherent and contextually plausible. Notably, the model preserves the professional tone and legal intent of the complaint, indicating strong semantic retention despite extensive perturbations.

Robustness under implicit and sparse identifiers. Figure 15 presents a scenario where sensitive information is partially implicit or sparsely distributed, including usernames, social links, and contact information embedded within a narrative. The masking process consistently anonymizes these elements (e.g., "@JessicaL82.com" $\rightarrow$ "@*******82.com") without disrupting readability. In reconstruction, the model replaces them with semantically consistent alternatives (e.g., "@contacthub82.com"), maintaining the functional role of each entity. Importantly, the generated substitutes align with the surrounding context (e.g., professional email domains, plausible phone formats), demonstrating that PrivMask captures latent constraints of entity usage rather than performing naive token replacement.

Privacy preservation in high-risk legal contexts. The example in Figure 16 highlights a particularly sensitive scenario involving legal disputes and financial documents, where improper handling of personal data could lead to severe privacy risks. The masking stage successfully removes identifiable signals such as URLs, usernames, and personal names, while retaining sufficient context for downstream reasoning. During reconstruction, PrivMask generates fully synthetic yet contextually appropriate entities (e.g., "Daniel Carter" and "https://financeinsight.org"), effectively breaking the linkage to original individuals. At the same time, the logical flow of the legal inquiry—covering court orders, compliance concerns, and privacy risks—remains intact, indicating that the model does not sacrifice critical semantic information.

H.6.3 Case Study on SPY-Medical

Figure 17, Figure 18 and Figure 19 illustrate representative examples from the SPY-Medical dataset, highlighting the challenges of anonymizing medical narratives that tightly couple personal identifiers with clinically meaningful content. Compared with other domains, medical texts require stronger consistency between entities (e.g., patient–doctor–institution relations) while preserving symptom descriptions and care context.

In Figure 17, the original text contains a dense mixture of PII, including addresses, practitioner names, emails, phone numbers, insurance providers, and institutional references. The masking stage effectively removes these identifiers while retaining the clinical narrative (symptoms, prior

Original Text	Masked Text	Reconstruction Text
I'm reaching out for some legal guidance regarding a peculiar situation that's been ongoing at my workplace for a few months now. As a dentist, I take pride in providing top-notch care to my patients, but I've started to feel a growing sense of unease regarding the handling of patient records by our office administrator, who I'll be meeting with at 443i Oakwood Hills Parkway, Wichita Falls, Tx 76301 for an urgent discussion soon. It's come to my attention that they've been consistently delaying the digital scanning of patient files, resulting in a massive backlog of paperwork that's now threatening to engulf our entire filing system. I've tried addressing this issue with the administrator on multiple occasions, but they seem reluctant to take responsibility or make tangible changes to their workflow, which is quite frustrating for someone like me who's deeply committed to patient care. I've tried escalating this issue to our practice manager, Sophia Rodriguez, who I've been trying to get on a call with via my personal $< PHONE_NUM >$, but they seem reluctant to get involved, citing "more pressing concerns" as the reason. I'm starting to feel a bit frustrated and isolated in this situation, as it seems like nobody is taking my concerns seriously. That's why I'm seeking some legal advice on the best course of action to take. You can check my $< URL >$ for more information about my background as a dentist, where I've worked with numerous clients, which is also linked to my $< USERNAME >$. Can someone please enlighten me on the potential legal implications of having unsecured patient records lying around? I'd also appreciate any guidance on how I can effectively advocate for change within my workplace without jeopardizing my professional relationships. Please feel free to reach out to me at $< EMAIL >$ if you're willing to offer any assistance. Additionally, I'd like to reach out to a patient advocate, Emily Mitchell at emily.mitcheli@healthadvocate.org, or her colleague, James Parker, who can be contacted at 317-455-3435 ext. 555. They have a resourceful website at https://healthadvocate.com that provides guidance on protecting patient records. I've heard that the state dental association, DAA at daa.org has similar resources, and their executive director, Michael Lee, is available at 512-367-1901. Thank you so much for taking the time to read through this and offer your expertise—I'm truly grateful for any help I can get, especially as this situation is affecting my professional reputation and overall well-being.	I'm reaching out for some legal guidance regarding a peculiar situation that's been ongoing at my workplace for a few months now. As a dentist, I take pride in providing top-notch care to my patients, but I've started to feel a growing sense of unease regarding the handling of patient records by our office administrator, who I'll be meeting with at **** ***wood Hills Parkway, Wichita Falls, Tx **301 for an urgent discussion soon. It's come to my attention that they've been consistently delaying the digital scanning of patient files, resulting in a massive backlog of paperwork that's now threatening to engulf our entire filing system. I've tried addressing this issue with the administrator on multiple occasions, but they seem reluctant to take responsibility or make tangible changes to their workflow, which is quite frustrating for someone like me who's deeply committed to patient care. I've tried escalating this issue to our practice manager, ****** Rodriguez, who I've been trying to get on a call with via my personal $< PHONE_NUM >$, but they seem reluctant to get involved, citing "more pressing concerns" as the reason. I'm starting to feel a bit frustrated and isolated in this situation, as it seems like nobody is taking my concerns seriously. That's why I'm seeking some legal advice on the best course of action to take. You can check my $< URL >$ for more information about my background as a dentist, where I've worked with numerous clients, which is also linked to my $< USERNAME >$. Can someone please enlighten me on the potential legal implications of having unsecured patient records lying around? I'd also appreciate any guidance on how I can effectively advocate for change within my workplace without jeopardizing my professional relationships. Please feel free to reach out to me at $< EMAIL >$ if you're willing to offer any assistance. Additionally, I'd like to reach out to a patient advocate, ***ly Mitchell at ***ly.*@healthadvocate.org, or her colleague, ***** Parker, who can be contacted at ***-***-3435 ext. 555. They have a resourceful website at https://******advocate.com that provides guidance on protecting patient records. I've heard that the state dental association, *** at ***.org has similar resources, and their executive director, *****el Lee, is available at ***-367-1901. Thank you so much for taking the time to read through this and offer your expertise—I'm truly grateful for any help I can get, especially as this situation is affecting my professional reputation and overall well-being.	I'm reaching out for some legal guidance regarding a peculiar situation that's been ongoing at my workplace for a few months now. As a dentist, I take pride in providing top-notch care to my patients, but I've started to feel a growing sense of unease regarding the handling of patient records by our office administrator, who I'll be meeting with at 4825 Crestwood Hills Parkway, Wichita Falls, Tx 11301 for an urgent discussion soon. It's come to my attention that they've been consistently delaying the digital scanning of patient files, resulting in a massive backlog of paperwork that's now threatening to engulf our entire filing system. I've tried addressing this issue with the administrator on multiple occasions, but they seem reluctant to take responsibility or make tangible changes to their workflow, which is quite frustrating for someone like me who's deeply committed to patient care. I've tried escalating this issue to our practice manager, Marisol Rodriguez, who I've been trying to get on a call with via my personal (940)-582-3479, but they seem reluctant to get involved, citing "more pressing concerns" as the reason. I'm starting to feel a bit frustrated and isolated in this situation, as it seems like nobody is taking my concerns seriously. That's why I'm seeking some legal advice on the best course of action to take. You can check my https://drjamesmillerdds.com for more information about my background as a dentist, where I've worked with numerous clients, which is also linked to my @drmiller_tx. Can someone please enlighten me on the potential legal implications of having unsecured patient records lying around? I'd also appreciate any guidance on how I can effectively advocate for change within my workplace without jeopardizing my professional relationships. Please feel free to reach out to me at j.miller@wfdentalgroup.org if you're willing to offer any assistance. I'd like to reach out to a patient advocate, Molly Mitchell at molly.c@healthadvocate.org, or her colleague, Daniel Parker, who can be contacted at 728-564-3435 ext. 555. They have a resourceful website at https://careadvocate.com that provides guidance on protecting patient records. I've heard that the state dental association, ADA at dental.org has similar resources, and their executive director, Daniel Lee, is available at 812-367-1901. Thank you so much for taking the time to read through this and offer your expertise—I'm truly grateful for any help I can get, especially as this situation is affecting my professional reputation and overall well-being.

Figure 14: An example from the SPY-Legal dataset showing the original, masked, and reconstruction snippets.

treatment, and consultation flow). In reconstruction, the model generates realistic surrogates such as plausible addresses, clinician identities, and contact details. Importantly, cross-entity consistency is preserved (e.g., specialist referrals and organizational roles), and the resulting text remains coherent and medically interpretable.

Figure 18 focuses on a longitudinal consultation scenario involving IBS, where structured identifiers (e.g., insurance ID, contact channels, and clinic location) are embedded within a self-reported symptom timeline. The masking process correctly anonymizes both explicit identifiers and semi-structured fields. The reconstruction step produces semantically aligned replacements that maintain logical relationships among entities (doctor–assistant–clinic) and preserve the temporal and diag-

Original Text	Masked Text	Reconstruction Text
I am $< NAME >$, currently working on a project for a local museum, where we are about to launch a new exhibit featuring artifacts from an ancient civilization. As part of our agreement with the exhibit's benefactors, we are required to provide a guarantee that the artifacts will be displayed and handled with the utmost care and respect. However, during the installation process, one of our team members accidentally damaged one of the artifacts, which has left us in a bit of a predicament. We've tried to repair the damage as best we could in-house, but we're not confident that our repairs meet the benefactors' standards, and we're worried that if the benefactors find out, they may withdraw their support for the exhibit. I've consulted with our in-house legal team at @JessicaL82.com, who have advised us to document everything and prepare for the worst-case scenario. However, I was wondering if anyone with more experience in this area could offer some guidance on the best course of action to take. Specifically, I'd like to know if we have any legal obligations to disclose the damage to the benefactors, and what our options are for addressing the situation without compromising the exhibit's integrity or the museum's reputation. You can reach me at $< EMAIL >$ or call me at $< PHONE_NUM >$ for any additional information or questions. I'd be eternally grateful for any guidance or insight that could help us navigate this tricky situation. Please feel free to share your thoughts and opinions at https://facebook.com/GreyHound2001. I'd appreciate any advice from experienced professionals like Dr. Sophia Patel, sophia.patel@lawfirm.com, 415-456-7890, or Mr. Michael Brown, mbrown@consultant.com, 202-555-1234, with a track record of handling sensitive issues similar to ours. You can also reach out to me directly at $< ADDRESS >$, and I will make sure to respond promptly.	I am $< NAME >$, currently working on a project for a local museum, where we are about to launch a new exhibit featuring artifacts from an ancient civilization. As part of our agreement with the exhibit's benefactors, we are required to provide a guarantee that the artifacts will be displayed and handled with the utmost care and respect. However, during the installation process, one of our team members accidentally damaged one of the artifacts, which has left us in a bit of a predicament. We've tried to repair the damage as best we could in-house, but we're not confident that our repairs meet the benefactors' standards, and we're worried that if the benefactors find out, they may withdraw their support for the exhibit. I've consulted with our in-house legal team at @*******82.com, who have advised us to document everything and prepare for the worst-case scenario. However, I was wondering if anyone with more experience in this area could offer some guidance on the best course of action to take. Specifically, I'd like to know if we have any legal obligations to disclose the damage to the benefactors, and what our options are for addressing the situation without compromising the exhibit's integrity or the museum's reputation. You can reach me at $< EMAIL >$ or call me at $< PHONE_NUM >$ for any additional information or questions. I'd be eternally grateful for any guidance or insight that could help us navigate this tricky situation. Please feel free to share your thoughts and opinions at https://facebook.com/*******2001. I'd appreciate any advice from experienced professionals like Dr. Sophia *****, **********@lawfirm.com, 415-***-****, or Mr. *****el Brown, ******@consultant.com, 202-***-****, with a track record of handling sensitive issues similar to ours. You can also reach out to me directly at $< ADDRESS >$, and I will make sure to respond promptly.	I am David Carter, currently working on a project for a local museum, where we are about to launch a new exhibit featuring artifacts from an ancient civilization. As part of our agreement with the exhibit's benefactors, we are required to provide a guarantee that the artifacts will be displayed and handled with the utmost care and respect. However, during the installation process, one of our team members accidentally damaged one of the artifacts, which has left us in a bit of a predicament. We've tried to repair the damage as best we could in-house, but we're not confident that our repairs meet the benefactors' standards, and we're worried that if the benefactors find out, they may withdraw their support for the exhibit. I've consulted with our in-house legal team at @contacthub82.com, who have advised us to document everything and prepare for the worst-case scenario. However, I was wondering if anyone with more experience in this area could offer some guidance on the best course of action to take. Specifically, I'd like to know if we have any legal obligations to disclose the damage to the benefactors, and what our options are for addressing the situation without compromising the exhibit's integrity or the museum's reputation. You can reach me at d.carter@citymuseum.org or call me at (617)-482-1934 for any additional information or questions. I'd be eternally grateful for any guidance or insight that could help us navigate this tricky situation. Please feel free to share your thoughts and opinions at https://facebook.com/community2001. I'd appreciate any advice from experienced professionals like Dr. Sophia Turner, advisor.team@lawfirm.com, 415-682-9134, or Mr. Daniel Brown, contact@consultant.com, 202-547-8821, with a track record of handling sensitive issues similar to ours. You can also reach out to me directly at 128 Museum Avenue, Boston, MA 02114, and I will make sure to respond promptly.

Figure 15: An example from the SPY-Legal dataset showing the original, masked, and reconstruction snippets.

nostic context. This demonstrates robustness in handling mixed structured–unstructured medical information.

Finally, Figure 19 presents a more narrative-driven case involving symptom descriptions, environmental triggers, historical treatment, and multi-party interactions (colleague, physician, and insurer). Despite the increased complexity, the masking stage consistently removes sensitive attributes across all entity types. The reconstructed text not only replaces these attributes with plausible alternatives but also preserves subtle clinical semantics such as symptom progression and contextual factors (e.g., anxiety, humidity). This indicates that the model can maintain fine-grained meaning while ensuring privacy.

I Training Details

In the privacy recognition stage, the BERT-base-uncased model is employed as the named entity recognition (NER) framework. The model is jointly fine-tuned on three datasets to enhance cross-domain privacy entity detection capability. The training configuration is as follows: batch size = 32, learning rate = 5e-5, and training epochs = 3. The AdamW optimizer is used with a linear learning rate decay and early stopping strategy to prevent overfitting. A class-weighted cross-entropy loss function is adopted to address the imbalance among different privacy entity types.

Original Text	Masked Text	Reconstruction Text
I'm currently seeking some guidance on a complex legal situation and I'm hoping someone can shed some light on my predicament. Here's the context: I've recently become the authorized representative for a family member who's facing a rather intricate lawsuit involving a disputed property ownership claim. As I'm not a legal professional myself, I'm doing my best to navigate this complicated process, but I'm struggling to understand the implications of a recent court ruling, and I'm worried about its impact on my family's financial security, especially considering the uncertain economic climate, which I follow closely on https://financeexpert.org. The ruling in question essentially ordered my family member to produce a range of financial documents, including tax returns and bank statements, within a tight deadline. I reached out to my trusted friend, a legal expert with a strong online presence at expertlawyer22, who cautioned me about the potential risks involved. Now, I completely understand the importance of complying with this order, but I'm concerned that providing these documents could potentially compromise my family member's privacy and put them at risk of identity theft. As $< NAME >$, I'm deeply worried about my own security as well, having previously experienced identity theft, and I've had to work closely with Julia Wilson during that difficult period. I've also consulted with a legal aid organization, which suggested that we may be able to negotiate an extension on the deadline or possibly even quash the subpoena altogether. However, I'm not confident that I fully grasp the legal nuances involved and would greatly appreciate any advice or guidance from someone with expertise in this area. Could someone please offer some insight into my options and potential risks associated with complying with this court order? If you're able to provide guidance, I'd be grateful if you could contact me directly at $< EMAIL >$, and I'll be happy to set up a call at your convenience. I'm particularly interested in understanding what measures I can take to safeguard my family member's privacy while still fulfilling our legal obligations, and I hope that someone with expertise in this area can provide valuable guidance. Any guidance or recommendations would be greatly appreciated. Additionally, if you could share your experience or advice via phone at $< PHONE_NUM >$, it would be helpful. Thanks in advance for your time and consideration.	I'm currently seeking some guidance on a complex legal situation and I'm hoping someone can shed some light on my predicament. Here's the context: I've recently become the authorized representative for a family member who's facing a rather intricate lawsuit involving a disputed property ownership claim. As I'm not a legal professional myself, I'm doing my best to navigate this complicated process, but I'm struggling to understand the implications of a recent court ruling, and I'm worried about its impact on my family's financial security, especially considering the uncertain economic climate, which I follow closely on https://finance******.org. The ruling in question essentially ordered my family member to produce a range of financial documents, including tax returns and bank statements, within a tight deadline. I reached out to my trusted friend, a legal expert with a strong online presence at expert*****22, who cautioned me about the potential risks involved. Now, I completely understand the importance of complying with this order, but I'm concerned that providing these documents could potentially compromise my family member's privacy and put them at risk of identity theft. As $< NAME >$, I'm deeply worried about my own security as well, having previously experienced identity theft, and I've had to work closely with Julia W***** during that difficult period. I've also consulted with a legal aid organization, which suggested that we may be able to negotiate an extension on the deadline or possibly even quash the subpoena altogether. However, I'm not confident that I fully grasp the legal nuances involved and would greatly appreciate any advice or guidance from someone with expertise in this area. Could someone please offer some insight into my options and potential risks associated with complying with this court order? If you're able to provide guidance, I'd be grateful if you could contact me directly at $< EMAIL >$, and I'll be happy to set up a call at your convenience. I'm particularly interested in understanding what measures I can take to safeguard my family member's privacy while still fulfilling our legal obligations, and I hope that someone with expertise in this area can provide valuable guidance. Any guidance or recommendations would be greatly appreciated. Additionally, if you could share your experience or advice via phone at $< PHONE_NUM >$, it would be helpful. Thanks in advance for your time and consideration.	I'm currently seeking some guidance on a complex legal situation and I'm hoping someone can shed some light on my predicament. Here's the context: I've recently become the authorized representative for a family member who's facing a rather intricate lawsuit involving a disputed property ownership claim. As I'm not a legal professional myself, I'm doing my best to navigate this complicated process, but I'm struggling to understand the implications of a recent court ruling, and I'm worried about its impact on my family's financial security, especially considering the uncertain economic climate, which I follow closely on https://financeinsight.org. The ruling in question essentially ordered my family member to produce a range of financial documents, including tax returns and bank statements, within a tight deadline. I reached out to my trusted friend, a legal expert with a strong online presence at expertalpha22, who cautioned me about the potential risks involved. Now, I completely understand the importance of complying with this order, but I'm concerned that providing these documents could potentially compromise my family member's privacy and put them at risk of identity theft. As Daniel Carter, I'm deeply worried about my own security as well, having previously experienced identity theft, and I've had to work closely with Julia Watson during that difficult period. I've also consulted with a legal aid organization, which suggested that we may be able to negotiate an extension on the deadline or possibly even quash the subpoena altogether. However, I'm not confident that I fully grasp the legal nuances involved and would greatly appreciate any advice or guidance from someone with expertise in this area. Could someone please offer some insight into my options and potential risks associated with complying with this court order? If you're able to provide guidance, I'd be grateful if you could contact me directly at daniel.carter@emailservice.com, and I'll be happy to set up a call at your convenience. I'm particularly interested in understanding what measures I can take to safeguard my family member's privacy while still fulfilling our legal obligations, and I hope that someone with expertise in this area can provide valuable guidance. Any guidance or recommendations would be greatly appreciated. Additionally, if you could share your experience or advice via phone at (312)-555-7821, it would be helpful. Thanks in advance for your time and consideration.

Figure 16: An example from the SPY-Legal dataset showing the original, masked, and reconstruction snippets.

Table 12 presents the privacy entity recognition performance of the BERT model across the three datasets. The results demonstrate that BERT accurately identifies various privacy entities across different corpus types, exhibiting strong robustness and generalization ability—particularly in medical texts. This ensures high-quality anonymized input for subsequent semantic reconstruction.

For the semantic reconstruction model, we adopt Flan-T5-base, which is trained to jointly optimize semantic consistency and privacy preservation. The model minimizes a weighted combination of generation loss and contrastive loss, ensuring that the reconstructed text maintains contextual coherence while effectively suppressing the regeneration of sensitive entities. The training setup uses a batch size of 32, a learning rate of 5×10^{-5}, and 3 epochs, with an early stopping strategy based on validation BERTScore-F1 to prevent overfitting. The total loss is defined

Original Text	Masked Text	Reconstruction Text
I've been experiencing some discomfort in my abdominal region for the past few days particularly after meals at my $< ADDRESS >$. The pain is dull and persists for several hours. I've noticed some bloating and gas too. I've tried over-the-counter antacids, which provided temporary relief. However, the symptoms persist. I signed up on a health forum with the username $< USERNAME >$ and tried gathering information on similar issues. I've had issues with heartburn and acid reflux in the past, which were treated with prescription medication. I've been taking the medication as prescribed, but I'm not sure if it's still effective. I've been speaking to my doctor, Dr. Smith, who has asked me to provide more details, which I've sent to him via $< EMAIL >$. Dr. Rachel Lee has asked her colleague, Dr. Michael Brown, to review my case. He can be reached at mlbrown@umed.net. I've also been in touch with a nutritionist, Emily Wilson, who has an office at 4567 Oak Street, Anytown, CA 91234. You can visit her website at wilsonwellness.com for more information on her services. Meanwhile, my health insurance provider, MedCarePlus, has assigned a case manager, Karen Davis, who can be reached at (877) 345-1111. Could you please advise on the next steps to diagnose and treat this condition? You can reach me at $< EMAIL >$ if you need any additional information. Are there any additional tests or examinations required? Are there any changes I can make to my diet or lifestyle to alleviate the symptoms? I'd appreciate any guidance on how to manage this condition in the long term. I've searched online and came across $< URL >$ which provided some useful information, but I'd like a more personalized approach. I've also sent a notification to my personal health assistant via $< PHONE_NUMBER >$ asking for her input on the matter. Additionally, I've received a response from the American Gastroenterological Association (AGA) at aga@gastrowellness.org, who suggested I consult with a specialist, Dr. Sophia Patel, who has an office at 123 Main Street, San Francisco, CA 94105, and can be reached at spatel@sfendo.com or (415) 555-1234.	I've been experiencing some discomfort in my abdominal region for the past few days, particularly after meals at my $< ADDRESS >$. The pain is dull and persists for several hours. I've noticed some bloating and gas too. I've tried over-the-counter antacids, which provided temporary relief. However, the symptoms persist. I signed up on a health forum with the username $< USERNAME >$ and tried gathering information on similar issues. I've had issues with heartburn and acid reflux in the past, which were treated with prescription medication. I've been taking the medication as prescribed, but I'm not sure if it's still effective. I've been speaking to my doctor, Dr. *****, who has asked me to provide more details, which I've sent to him via $< EMAIL >$. Dr. ****el Lee has asked her colleague, Dr. ******* Brown, to review my case. He can be reached at ******@umed.net. I've also been in touch with a nutritionist, **ily Wilson, who has an office at **** Oak Street, Anytown, CA *****. You can visit her website at ***********.com for more information on her services. Meanwhile, my health insurance provider, **********, has assigned a case manager, ***** *****, who can be reached at (***) ***-1111. Could you please advise on the next steps to diagnose and treat this condition? You can reach me at $< EMAIL >$ if you need any additional information. Are there any additional tests or examinations required? Are there any changes I can make to my diet or lifestyle to alleviate the symptoms? I'd appreciate any guidance on how to manage this condition in the long term. I've searched online and came across $< URL >$ which provided some useful information, but I'd like a more personalized approach. I've also sent a notification to my personal health assistant via $< PHONE_NUM >$ asking for her input on the matter. Additionally, I've received a response from the American Gastroenterological Association (AGA) at aga@gastrowellness.org, who suggested I consult with a specialist, Dr. ******, who has an office at *** **** Street, San Francisco, CA ***05, and can be reached at ******@sfendo.com or ***** 555-1234.	I've been experiencing some discomfort in my abdominal region for the past few days, particularly after meals at my 1240 Pinecrest Avenue, Sunnyvale, CA 94087. The pain is dull and persists for several hours. I've noticed some bloating and gas too. I've tried over-the-counter antacids, which provided temporary relief. However, the symptoms persist. I signed up on a health forum with the username healing_journey89 and tried gathering information on similar issues. I've had issues with heartburn and acid reflux in the past, which were treated with prescription medication. I've been taking the medication as prescribed, but I'm not sure if it's still effective. I've been speaking to my doctor, Dr. Mason, who has asked me to provide more details, which I've sent to him via d.lith@baycareclinic.org. Dr. Daniel Lee has asked her colleague, Dr. Charles Brown, to review my case. He can be reached at jbrown@umed.net. I've also been in touch with a nutritionist, Laily Wilson, who has an office at 7890 Pine Street, Anytown, CA 91876. You can visit her website at healthhubcenter.com for more information on her services. Meanwhile, my health insurance provider, CareWellPro, has assigned a case manager, Laura Bennett, who can be reached at (855) 672-1111. Could you please advise on the next steps to diagnose and treat this condition? You can reach me at patient.contact@outlook.com if you need any additional information. Are there any additional tests or examinations required? Are there any changes I can make to my diet or lifestyle to alleviate the symptoms? I'd appreciate any guidance on how to manage this condition in the long term. I've searched online and came across https://digestivehealthportal.org/articles/acid-reflux-care which provided some useful information, but I'd like a more personalized approach. I've also sent a notification to my personal health assistant via (408) 555-1234 asking for her input on the matter. Additionally, I've received a response from the American Gastroenterological Association (AGA) at info@agrowellness.org, who suggested I consult with a specialist, Dr. Sneha Patel, who has an office at 123 Main Street, San Francisco, CA ###05, and can be reached at s#####@sfendo.com or (###) 555-1234.

Figure 17: An example from the SPY-Medical dataset showing the original, masked, and reconstruction snippets.

Table 12: Performance of Privacy Entity Recognition.

Dataset	Precision	Recall	F1-score
PII-Masking-300k-sub	94.7%	95.5%	95.2%
SPY-Legal	95.0%	96.4%	95.7%
SPY-Medical	96.1%	93.6%	94.8%

as $\mathcal{L}total = \mathcal{L}gen + 0.5\mathcal{L}_{InfoNCE}$, where the generation and contrastive loss weights are set to 1.0 and 0.5, respectively. For the InfoNCE objective, the temperature parameter is set to $\gamma = 0.07$. These values are selected through validation experiments and fixed across all datasets to ensure consistent training conditions across domains.

For the contrastive objective, the representation $\mathbf{z}$ is obtained by mean pooling the final-layer encoder hidden states over non-padding tokens. The positive representation $\mathbf{z}^+$ is obtained from the

Original Text	Masked Text	Reconstruction Text
I'm seeking an online consultation as $< EMAIL >$, I've been experiencing persistent abdominal discomfort and bloating over the past three weeks. The symptoms are worse after eating, and I've noticed a change in my bowel movements - they've become more frequent and urgent. I've also experienced some nausea, but no fever or vomiting. I've tried to manage my symptoms with over-the-counter antacids and a fibre supplement, but they don't seem to be providing much relief. I have a history of IBS, and I was diagnosed with it about five years ago, during a visit to my primary care physician whose address is $< ADDRESS >$. However, my symptoms have been well-controlled with diet and stress management, including meditation sessions on $< URL >$, until now. I'm worried that something else might be underlying this flare-up, that's why I started to document my symptoms on a $< USERNAME >$ shared spreadsheet, where I post regular health updates. Could you please advise on the best course of action to take? Should I continue with my current treatment, or would you recommend further testing or a different medication, which I would purchase with my insurance policy number $< ID_NUM >$? Dr. Emma Wilson could be available to discuss the case with you. You can reach her at evwilson@gmail.com, or on her office phone at 456-734-9876. Alternatively, her assistant, Mr. Ryan Stevens, can be contacted at rstevens72@outlook.com or dial 213-456-1234. Their office is located at 1234, Oak Street, Las Vegas, NV 89101. I'd also appreciate any guidance on how to manage my symptoms in the short-term while we work together to determine the cause of this flare-up. In the meantime, I can be reached at $< PHONE_NUM >$. Additionally, I'd like to know if there are any lifestyle changes I can make to prevent future episodes, including adjusting my daily routine, which is documented on $< URL >$. I'm open to any suggestions or recommendations you may have, and would love to discuss it further over the phone, especially with Dr. Sophia Patel, whose profile can be accessed at https://doctors.healthcare.com/sophiapatelmd. I'd appreciate your help.	I'm seeking an online consultation as $< EMAIL >$, I've been experiencing persistent abdominal discomfort and bloating over the past three weeks. The symptoms are worse after eating, and I've noticed a change in my bowel movements - they've become more frequent and urgent. I've also experienced some nausea, but no fever or vomiting. I've tried to manage my symptoms with over-the-counter antacids and a fibre supplement, but they don't seem to be providing much relief. I have a history of IBS, and I was diagnosed with it about five years ago, during a visit to my primary care physician whose address is $< ADDRESS >$. However, my symptoms have been well-controlled with diet and stress management, including meditation sessions on $< URL >$, until now. I'm worried that something else might be underlying this flare-up, that's why I started to document my symptoms on a $< USERNAME >$ shared spreadsheet, where I post regular health updates. Could you please advise on the best course of action to take? Should I continue with my current treatment, or would you recommend further testing or a different medication, which I would purchase with my insurance policy number $< ID_NUM >$? Dr. Emma W****** could be available to discuss the case with you. You can reach her at evw******@gmail.com, or on her office phone at ***-***-9876. Alternatively, her assistant, Mr. Ryan S******, can be contacted at rstevens**@outlook.com or dial ***-***-1234. Their office is located at ****, Oak Street, Las Vegas, ** *****. I'd also appreciate any guidance on how to manage my symptoms in the short-term while we work together to determine the cause of this flare-up. In the meantime, I can be reached at $< PHONE_NUM >$. Additionally, I'd like to know if there are any lifestyle changes I can make to prevent future episodes, including adjusting my daily routine, which is documented on $< URL >$. I'm open to any suggestions or recommendations you may have, and would love to discuss it further over the phone, especially with Dr. Sophia P****, whose profile can be accessed at https://doctors.*********.com/sophiap******. I'd appreciate your help.	I'm seeking an online consultation as patient.consult@healthmail.com, I've been experiencing persistent abdominal discomfort and bloating over the past three weeks. The symptoms are worse after eating, and I've noticed a change in my bowel movements - they've become more frequent and urgent. I've also experienced some nausea, but no fever or vomiting. I've tried to manage my symptoms with over-the-counter antacids and a fibre supplement, but they don't seem to be providing much relief. I have a history of IBS, and I was diagnosed with it about five years ago, during a visit to my primary care physician whose address is 5678, Maple Avenue, Dallas, TX 75201. However, my symptoms have been well-controlled with diet and stress management, including meditation sessions on https://calmhealthhub.com, until now. I'm worried that something else might be underlying this flare-up, that's why I started to document my symptoms on a healthlog_user23 shared spreadsheet, where I post regular health updates. Could you please advise on the best course of action to take? Should I continue with my current treatment, or would you recommend further testing or a different medication, which I would purchase with my insurance policy number ZX-48291-PL? Dr. Emma Walker could be available to discuss the case with you. You can reach her at evwalker@gmail.com, or on her office phone at (972)-845-9876. Alternatively, her assistant, Mr. Ryan Sanders, can be contacted at rstevens11@outlook.com or dial (214)-555-1234. Their office is located at Unit, Oak Street, Las Vegas, NV 89101. I'd also appreciate any guidance on how to manage my symptoms in the short-term while we work together to determine the cause of this flare-up. In the meantime, I can be reached at (469)-882-1934. Additionally, I'd like to know if there are any lifestyle changes I can make to prevent future episodes, including adjusting my daily routine, which is documented on https://wellnessroutine.net. I'm open to any suggestions or recommendations you may have, and would love to discuss it further over the phone, especially with Dr. Sophia Perry, whose profile can be accessed at https://doctors.healthhub.com/sophiaperrymd. I'd appreciate your help.

Figure 18: An example from the SPY-Medical dataset showing the original, masked, and reconstruction snippets.

corresponding original, unmasked sentence, while negative representations $\mathbf{z}_i^-$ are constructed using other original sentences within the same mini-batch. Thus, each sample uses the remaining $N = B - 1$ instances in a batch as negative candidates, where $B = 32$ denotes the batch size. No unmasked sensitive entity is directly used as a negative target, preventing the contrastive objective from explicitly exposing private content during training. The InfoNCE loss is therefore computed using one positive pair and 31 in-batch negative pairs for each training instance.

During training and validation, BERTScore is employed as the primary evaluation metric, measuring semantic alignment between the reconstructed and original texts through three components: Precision, Recall, and F1-score. Table 13 summarizes the semantic reconstruction performance of the Flan-T5 model on the three datasets.

Overall, the results indicate that Flan-T5 exhibits strong semantic reconstruction performance across both general and domain-specific corpora (legal and medical), maintaining BERTScore-F1 values consistently above 97%. The relatively small variation in Recall across datasets indicates stable

Original Text	Masked Text	Reconstruction Text
I am $< NAME >$, and I have been experiencing a peculiar sensation in my chest, which I can only describe as a dull ache, like a faint echo of a distant drumbeat. It has been persisting for the past fortnight, and I've noticed it tends to intensify when I'm in environments with high humidity or during periods of heightened anxiety. My medical history is largely unremarkable, save for a bout of pneumonia I contracted during a research expedition to the British Museum's archives several years ago, where I occasionally met with my colleague Olivia Nelson, reachable at olivia.nelson@example.com and 234-567-5678, who lives at 1436 F Street, Washington D.C.. I was treated with a course of antibiotics and made a full recovery, which I kept track of on my personal portal accessible via $< URL >$. In an effort to alleviate this present discomfort, I've been taking a daily dose of omega-3 supplements and practicing gentle yoga to regulate my breathing, often streaming online classes through $< URL >$. While these measures have provided some relief, the sensation persists, and I fear it may be indicative of a more serious underlying issue, which I'd like to discuss with Dr. Sophia Patel, who can be reached at 987-654-3210 ext. 456, or at sophia.patel@example.com, and is located at Regional Medical Center, 321 1st Ave, New York, NY 10001. I would appreciate it if we could discuss my symptoms further over the phone, I am available at $< PHONE_NUM >$. Could you please advise on the possibility of an underlying cardiac condition, and recommend a course of action to further investigate and address this issue? I've been tracking my symptoms on my personal health portal, accessible via $< URL >$, and I can share the username $< USERNAME >$ and password with you for your review. Additionally, I would appreciate any guidance on the potential correlation between environmental factors and this symptom, as I am keen to understand the root cause of this affliction. I'll be keeping an eye on my email $< EMAIL >$ for your response. Please note that all of my medical records are linked to my personal identifier $< ID_NUM >$, which I've had since my first medical visit to the University Health Services. If you need any additional information, you can contact my insurance provider, HealthCare Inc, at healthcare_inc@example.com, 123-456-7890, or at their office located at 4567 Oak St, Chicago, IL 60611.	I am $< NAME >$, and I have been experiencing a peculiar sensation in my chest, which I can only describe as a dull ache, like a faint echo of a distant drumbeat. It has been persisting for the past fortnight, and I've noticed it tends to intensify when I'm in environments with high humidity or during periods of heightened anxiety. My medical history is largely unremarkable, save for a bout of pneumonia I contracted during a research expedition to the British Museum's archives several years ago, where I occasionally met with my colleague Olivia ******, reachable at olivia.******@example.com and ***-***-5678, who lives at **** * Street, Washington D.C.. I was treated with a course of antibiotics and made a full recovery, which I kept track of on my personal portal accessible via $< URL >$. In an effort to alleviate this present discomfort, I've been taking a daily dose of omega-3 supplements and practicing gentle yoga to regulate my breathing, often streaming online classes through $< URL >$. While these measures have provided some relief, the sensation persists, and I fear it may be indicative of a more serious underlying issue, which I'd like to discuss with Dr. ****** ******, who can be reached at ***-654-3210 ext. 456, or at **********@example.com, and is located at ******** **********, *** 1st Ave, New York, NY 10001. I would appreciate it if we could discuss my symptoms further over the phone, I am available at $< PHONE_NUM >$. Could you please advise on the possibility of an underlying cardiac condition, and recommend a course of action to further investigate and address this issue? I've been tracking my symptoms on my personal health portal, accessible via $< URL >$, and I can share the username $< USERNAME >$ and password with you for your review. Additionally, I would appreciate any guidance on the potential correlation between environmental factors and this symptom, as I am keen to understand the root cause of this affliction. I'll be keeping an eye on my email $< EMAIL >$ for your response. Please note that all of my medical records are linked to my personal identifier $< ID_NUM >$, which I've had since my first medical visit to the University Health Services. If you need any additional information, you can contact my insurance provider, ********** Inc, at **********_inc@example.com, ***-***-7890, or at their office located at **** *** St, Chicago, IL *****.	I am Ethan Clarke, and I have been experiencing a peculiar sensation in my chest, which I can only describe as a dull ache, like a faint echo of a distant drumbeat. It has been persisting for the past fortnight, and I've noticed it tends to intensify when I'm in environments with high humidity or during periods of heightened anxiety. My medical history is largely unremarkable, save for a bout of pneumonia I contracted during a research expedition to the British Museum's archives several years ago, where I occasionally met with my colleague Olivia Carter, reachable at olivia.carter@example.com and (202)-555-5678, who lives at 1520 Pine Street, Washington, D.C.. I was treated with a course of antibiotics and made a full recovery, which I kept track of on my personal portal accessible via https://healthtrackerhub.com. In an effort to alleviate this present discomfort, I've been taking a daily dose of omega-3 supplements and practicing gentle yoga to regulate my breathing, often streaming online classes through https://yogawellnesslive.com. While these measures have provided some relief, the sensation persists, and I fear it may be indicative of a more serious underlying issue, which I'd like to discuss with Dr. Amelia Carter, who can be reached at 012-654-3210 ext. 456, or at amelia.carter@example.com, and is located at Central Health Clinic, 789 1st Ave, New York, NY 10001. I would appreciate it if we could discuss my symptoms further over the phone, I am available at (917)-555-1182. Could you please advise on the possibility of an underlying cardiac condition, and recommend a course of action to further investigate and address this issue? I've been tracking my symptoms on my personal health portal, accessible via https://symptomlog.net, and I can share the username ethan_healthlog and password with you for your review. Additionally, I would appreciate any guidance on the potential correlation between environmental factors and this symptom, as I am keen to understand the root cause of this affliction. I'll be keeping an eye on my email ethan.clarke@mailhub.com for your response. Please note that all of my medical records are linked to my personal identifier ID-59382-XZ, which I've had since my first medical visit to the University Health Services. If you need any additional information, you can contact my insurance provider, MedSecure Inc, at medscure_inc@example.com, (800)-555-7890, or at their office located at 7890 Pine St, Chicago, IL 10177.

Figure 19: An example from the SPY-Medical dataset showing the original, masked, and reconstruction snippets.

Table 13: Semantic Reconstruction Performance (BERTScore).

Dataset	Precision	Recall	F1-score
PII-Masking-300k-sub	97.1%	97.5%	97.3%
SPY-Legal	97.5%	97.6%	97.5%
SPY-Medical	97.3%	97.4%	97.4%

semantic reconstruction performance across different domains. These findings support the applicability of the proposed semantic-privacy weighted training strategy across diverse data domains.

Table 14: Categories of Sensitive Information and Corresponding Sensitivity Levels.

Category	Examples	Sensitivity Level	Rationale
Unique Identification	ID number, Passport number, Social Security Number (SSN)	High	Enables direct identification or identity theft; listed as highly identifiable by GDPR and similar regulations
Financial	Bank account, Credit card number, Transaction records	High	Directly causes financial loss; high compliance risk (PCI, GDPR)
Health/Medical	Diagnosis, Medical record number, Prescription, Medication history	High	Protected Health Information (PHI) under HIPAA; high privacy sensitivity
Biometric	Fingerprints, Facial features, Iris template, DNA	High	Immutable identifiers; misuse has serious consequences
Precise Geolocation	GPS coordinates, Real-time location data	High	Can track or locate individuals; re-identifiable when combined
Credentials & Keys	Password, API key, Private key, Session token	High	Can directly lead to account takeover or system intrusion
Contact	Phone number, Email address, Home telephone	Medium	Enables social engineering or identity inference when combined
Address	Full residential address, Street number	Medium	Enables individual localization; risk varies by combination
Employment	Position, Employer, Employee ID	Medium	Associated with occupational or social risk; raises re-identification potential
Demographics	Birth date, Age range, Gender, Ethnicity	Low/Medium	Low risk alone but identifiable when combined with others
Full Name	Real name, Nickname	Low	Common information, low standalone risk (higher in combination)
Device IDs	IMEI, MAC address, Device fingerprint	Medium	Used for tracking and device-level identification
Network Identifiers	Public IP, Port number	Medium	Locates users or networks; short-term identifiability
Behavioral/Preference	Browsing history, Clickstream, Purchase preference	Medium	Enables profiling and inference of sensitive attributes
Regulated Data	Tax ID, Social security record	High	Legally regulated with strict data protection requirements
Family/Relationship	Family members, Relationship chain	Medium	Reveals household structure; privacy implications
Legal/Criminal	Arrest record, Judgments, Fines	High/Medium	Significant personal impact; requires lawful handling
Education	School, Degree, Transcript	Low/Medium	Low sensitivity alone; certificate ID increases risk
Location History	Long-term mobility trajectory	High	Long-term trajectories enable re-identification
Metadata	Call duration, Session timestamp	Low/Medium	Enables side-channel inference
Transaction Behavior	Transaction amount, Merchant, Time	Medium/High	High sensitivity when linked to financial accounts
Genomic Data	DNA sequence, Gene variation	High	Permanent identifier; impacts individual and family
Visual Media	Identifiable photo or video	High	Enables face recognition or re-identification
Voice	Audio sample, Voiceprint	High	Enables biometric identification
Inferred Sensitive	Political views, Religion, Sexual orientation (inferred)	High (inferred)	Strictly regulated under GDPR due to inference sensitivity
Temporal Identifiers	One-time password (OTP)	High (short-term)	Directly usable for authentication
Facility Identifiers	Hospital ID, Surveillance location	Medium/High (contextual)	Sensitive when linked to individuals
Social Handles	Twitter, Weibo account	Medium	Used for tracking or social engineering
Research IDs	Participant ID, Experiment code	Low/Medium	Sensitive if reversibly traceable
Custom Identifiers	Customer ID, Internal code	Context-dependent	Sensitive if mappable to individuals

J Detailed Classification of Sensitive Information and Corresponding Sensitivity Levels

Table 14 provides a comprehensive classification of sensitive information types along with their corresponding sensitivity levels and underlying rationale.