Anti-adversarial Learning: Desensitizing Prompts for Large Language Models

Xuan Li, Zhe Yin, Xiaodong Gu, Beijun Shen*

School of Computer Science, Shanghai Jiao Tong University, China
{riken01, yin_zhe, xiaodong.gu, bjshen}@sjtu.edu.cn

Abstract

With the widespread use of LLMs, preserving privacy in user prompts has become crucial, as prompts risk exposing private and sensitive data to cloud LLMs. Conventional techniques like homomorphic encryption (HE), secure multi-party computation, and federated learning (FL) are not well-suited to this scenario due to the lack of control over user participation in remote model interactions. In this paper, we propose PromptObfus, a novel method for desensitizing LLM prompts. The core idea of PromptObfus is "anti-adversarial" learning. Unlike adversarial attacks that add imperceptible perturbations to mislead models, PromptObfus perturbs sensitive words to make them unrecognizable to humans while maintaining the model's original predictions. Specifically, PromptObfus frames prompt desensitization as a masked language modeling task, replacing privacy-sensitive terms with a `[MASK]` token. A desensitization model is utilized to generate candidate replacements for each masked position. These candidates are subsequently selected based on gradient feedback from a surrogate model, ensuring minimal disruption to task output. We demonstrate the effectiveness of our approach on three NLP tasks. Results show that PromptObfus effectively prevents privacy inference from remote LLMs while preserving task utility.

Code and Datasets —
https://github.com/riken01/PromptObfus

Introduction

The widespread adoption of large language models (LLMs) such as ChatGPT in various NLP tasks (Yao et al. 2024) has raised significant concerns regarding their inherent privacy risks. Due to the substantial computational resources required for local deployment, users often rely on cloud APIs provided by model vendors, which introduces potential vulnerabilities. Specifically, user-submitted prompts, the primary medium of interaction with LLMs, may inadvertently expose sensitive information, posing serious privacy threats.

Prompts often contain personally identifiable information (PII), including names, addresses, and occupational details, as illustrated in Figure 1. Without proper safeguards during processing, this sensitive data becomes vulnerable to malicious exploitation, leading to serious privacy breaches (Duan et al. 2024). Thus, developing robust privacy protection mechanisms for LLM prompts has become an urgent research priority.

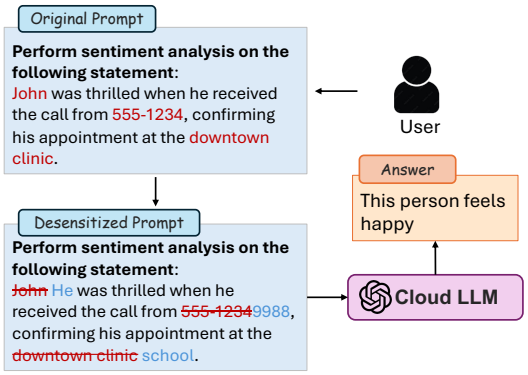

Figure 1: Illustration of prompt desensitization.

Conventional privacy-preserving techniques, such as Homomorphic Encryption (HE) (Gentry 2009), Secure Multi-Party Computation (MPC) (Yao 1982), and Federated Learning (FL) (McMahan et al. 2017), exhibit significant limitations when applied to prompts for LLMs, particularly in black-box settings where access to the model's internal architecture or training data is restricted. These methods often fail to simultaneously address the competing requirements of real-time performance, computational efficiency, and robust privacy protection.

Text obfuscation has emerged as a prevalent approach to safeguarding sensitive information in prompts (Miranda et al. 2025). For instance, techniques include injecting noise into word embeddings based on differential privacy to perturb sensitive data (Yue et al. 2021; Gao et al. 2025), clustering word vectors to render representations of sensitive terms indistinguishable (Zhou et al. 2023), and training models for data anonymization by detecting and removing PII entities (Chen et al. 2023; Frikha et al. 2025). However, these methods often struggle to achieve an optimal trade-off between privacy preservation and task utility (Zhang et al. 2025). Fur-

*The corresponding author.

thermore, approaches that rely on model training typically necessitate expert-annotated datasets, which are challenging to procure in practical applications.

In this paper, we propose PromptObfus, a portable and task-flexible method for desensitization of LLM prompts. Inspired by the work on generating adversarial examples (Alzantot et al. 2018), we introduce the concept of *anti-adversariality*, which inverts the adversarial objective: rather than crafting subtle perturbations to mislead models, we strategically alter sensitive words to make them unrecognizable to human interpreters while maintaining the model's original task performance. PromptObfus achieves desensitization by replacing words with semantically distinct yet task-consistent alternatives, thereby ensuring robust privacy protection without compromising the original functionality of the prompts. PromptObfus operates through the deployment of two small local models: a *desensitization model*, which replaces sensitive words with privacy-preserving alternatives, and a *surrogate model*, which emulates the task execution of the remote LLM to guide prompt selection. The pipeline consists of three critical steps: generating desensitized alternatives for privacy-sensitive words, assessing the task utility of the LLM, and selecting replacements that minimize performance degradation.

We evaluate PromptObfus on three NLP tasks: sentiment analysis, topic classification, and question answering. The results demonstrate that our approach establishes new state-of-the-art privacy protection, achieving a 62.70% reduction in implicit privacy inference attack success rates compared to existing high-accuracy baselines, while completely eliminating explicit inference attacks. Notably, our approach simultaneously preserves competitive task utility, yielding accuracy scores of 86.67%, 85.25%, and 96.0%, respectively.

Our contribution can be summarized as follows:

- We introduce the novel concept of **anti-adversariality**, a pioneering approach for desensitizing LLM prompts that ensures robust privacy protection without compromising task utility.
- We propose a new privacy-preserving word replacement algorithm, which integrates masked word prediction with LLM gradient surrogation to achieve optimal desensitization.
- We conduct extensive evaluations of our method across multiple NLP tasks, demonstrating its effectiveness in preserving privacy while preserving task utility.

Related Work

Privacy Protection for LLMs. Despite their widespread utility, LLMs raise critical privacy concerns (Mireshghallah et al. 2024). Current research addresses these through: (1) model protection via federated learning (Hu et al. 2024; Liu et al. 2025) and homomorphic encryption (Hao et al. 2022); (2) prompt security using encryption (Lin, Hua, and Zhang 2024) and noise-based obfuscation (Zhou et al. 2023; Gao et al. 2025); and (3) PII detection/removal techniques (Chen et al. 2023; Sun et al. 2024; Chowdhury et al. 2025). Hybrid input strategies mixing real and synthetic data further enhance privacy (Utpala, Hooker, and Chen 2023).

Automatic Prompt Engineering. Automatic prompt generation leverages AI to produce privacy-preserving prompts, offering superior performance compared to manual approaches (Zhou et al. 2022). Notable frameworks include APE (Yang et al. 2024), which iteratively refines prompts by selecting and resampling candidate prompts; APO (Zhou et al. 2022), employing gradient-inspired feedback optimization; and OPRO (Pryzant et al. 2023), utilizing LLMs as meta-optimizers for prompt improvement.

Text Adversary Generation. Adversarial training is a technique aimed at improving model robustness against malicious or deceptive inputs, widely applied in domains such as computer vision, NLP, and speech recognition. In this approach, models are systematically exposed to adversarial examples (Goodfellow et al. 2014), which are inputs subtly modified to induce significant changes in model outputs. Genetic algorithms are employed to generate semantically equivalent adversarial samples (Alzantot et al. 2018), selecting synonyms that maximize the likelihood of the target label. More recently, LLMs are utilized to produce adversarial samples (Wang et al. 2024).

In contrast to existing approaches, we propose an *anti-adversarial* method for the desensitization of LLM prompts, which ensures that model outputs remain consistent while rendering sensitive content imperceptible to human interpretation.

Methodology

Inspired by the principles of adversarial example generation (Alzantot et al. 2018), we conceptualize our approach as an *anti-adversarial* framework, wherein the objective is to obfuscate sensitive information while preserving the original behavior and predictive performance of the model.

Problem Statement

Consider an LLM $\Phi(y|x)$ with parameters Φ and a downstream task (*e.g.*, question answering) characterized by a parallel dataset $\mathcal{T} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, where x and y represent the input prompt and target output, respectively. We formulate the following privacy-preserving transformation problem: Given a set of privacy attributes $P = [p_1, \ldots, p_m]$ and an input $x = \{x_1, \ldots, x_n\}$, our goal is to derive a desensitized prompt $x' = \{x'_1, \ldots, x'_n\}$ that eliminates all P-attributes while preserving task utility. Formally:

$$\begin{aligned} \min_{x'=M(x|\lambda,k)} & \|s(\Phi(x'), y) - s(\Phi(x), y)\| \\ s.t. \quad & x'_i \notin P \quad \forall x'_i \in x' \end{aligned} \tag{1}$$

where $M(x|\lambda, k)$ denotes a desensitization mapping function, λ controls the candidate replacement set size for each sensitive term, and k modulates the confusion ratio. The task-specific metric $s : Y \times Y \to \mathbb{R}$ (*e.g.*, BLEU for QA) evaluates utility preservation.

Overview

Our approach is designed to optimize the desensitization function $M(x|\lambda, k)$ to preserve LLM output fidelity while

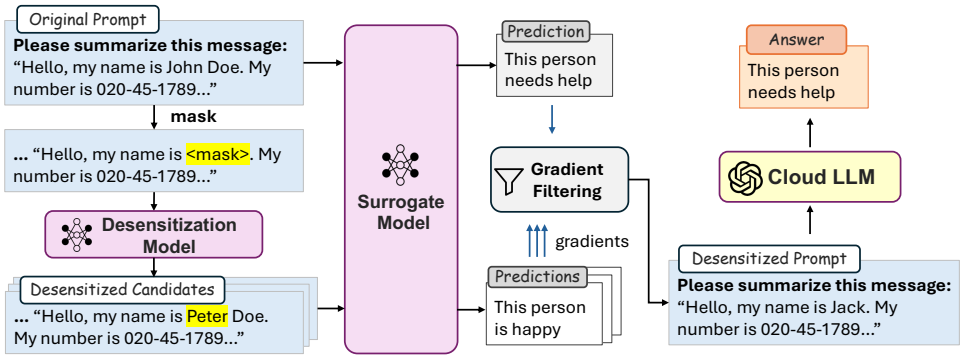

Figure 2: Overview of PromptObfus.

eliminating privacy risks. Figure 2 illustrates the overall architecture of PromptObfus. The pipeline consists of three steps: (1) detecting privacy attributes and generating candidate replacements using a dedicated desensitization model; (2) assessing utility preservation through a surrogate model by comparing with the original prompt's performance; and (3) performing gradient-based optimization to select the most suitable replacements from candidates, ultimately producing the final privacy-preserving prompt.

Predicting Candidate Desensitive Words

For each privacy-sensitive word in an input prompt, PromptObfus generates a set of candidate replacements through desensitization. This process can be formalized as a Masked Language Model (MLM) task, where privacy-sensitive words are substituted with a `mask` token. The desensitization model is utilized to predict precisely λ candidate desensitized replacements for each masked position. By leveraging pre-trained semantic representations, the model ensures all candidate replacements maintain contextual appropriateness relative to the surrounding text. This approach preserves textual coherence and prompt functionality while effectively concealing sensitive information through semantically valid substitutions.

We utilize spaCy's NER model (Explosion 2024) to detect explicit privacy attributes like person names, locations, and organizations. All identified privacy-sensitive words are uniformly replaced with MASK tokens. Beyond explicit attributes, we address potential implicit privacy risks through contextual analysis. Specifically, we mask rare words identified by their TF-IDF scores (Vats et al. 2024; Sparck Jones 1988), as these terms are statistically more likely to contain identifiable information. The top k highest-scoring terms are selected for masking.

Next, a pre-trained language model, referred to as the *desensitization model*, is utilized to generate potential replacement candidates for each masked token, as shown in Figure 3. This model can employ any pre-trained language architecture with MLM capability, such as RoBERTa.

To mitigate the risk of privacy leakage through synonyms or near-synonyms, the desensitized word set is further refined by assessing semantic similarity. This is achieved by computing the Euclidean distance between the word embeddings of each candidate w_i and the original word x_{original}:

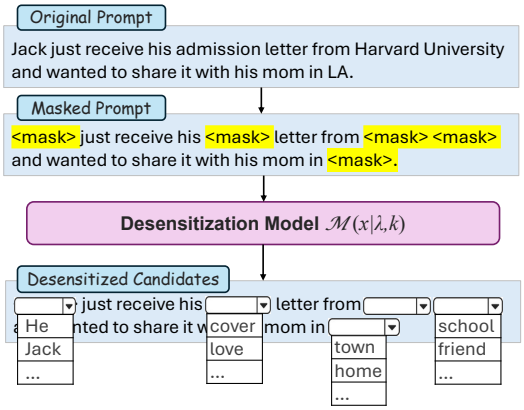

Figure 3: Predicting candidate desensitive words.

$$d(x_{\text{original}}, w_i) = \|\vec{x_{\text{original}}} - \vec{w_i}\| \quad (2)$$

where $\vec{x_{\text{original}}}$ and $\vec{w_i}$ are the word vectors, and $\|\cdot\|$ is the Euclidean norm. To this end, we apply an empirically determined distance threshold $\theta_{\text{dist}} = 0.95$. All candidates satisfying $d(x_{\text{original}}, w_i) \leq \theta_{\text{dist}}$ are considered semantically too similar and are consequently removed. The resulting filtered set W_{filtered} is defined as:

$$W_{\text{filtered}} = \{w_i \in W \mid d(x_{\text{original}}, w_i) > \theta_{\text{dist}}\} \quad (3)$$

Assessing Task Utility

To preserve task utility, we design a gradient-based selection criterion for desensitized words. Gradient magnitudes serve as indicators of input sensitivity: larger values suggest substantial semantic distortion from word replacement, while smaller values imply better semantic preservation with minimal output perturbation.

Since direct gradient acquisition from remote LLMs is infeasible, PromptObfus employs a smaller white-box surrogate model $\mathcal{M}_{surrogate}$ to approximate the target LLM's behavior, which captures how variations in the prompt affect the performance signal. This computationally efficient alternative enables both task evaluation and gradient computation while maintaining manageable resource requirements. PromptObfus supports two types of surrogate models:

1) Task-specific model: When adequate task-specific data $\mathcal{D} = \{(x, y)\}$ exists, a lightweight fine-tuned model provides precise, task-aware gradient estimates for prompt desensitization.

2) General model: For data-scarce scenarios, a moderately-sized pre-trained language model (still substantially smaller than target LLMs) serves as the surrogate. This variant produces less task-specific but more generalizable gradient approximations.

Gradient Filtering

PromptObfus utilizes gradient magnitudes from the surrogate model $\mathcal{M}_{surrogate}$ to assess desensitized candidates in $W_{filtered}$, selecting the word corresponding to the minimal gradient value.

For each candidate word $w \in W_{filtered}$, PromptObfus generates a modified prompt x' and computes its output gradient. Formally, the gradient magnitude is calculated as:

$$\Delta_i(w) = \left\| \frac{\partial \mathcal{L}(y, \mathcal{M}_{surrogate}(x'[i \leftarrow w]))}{\partial x'} \right\| \quad (4)$$

where i indicates the target word position, $\Delta_i(w)$ captures the gradient sensitivity, and $\mathcal{L}$ represents the task loss function. Through iterative evaluation, the optimal replacement w^* is selected via:

$$w^* = \arg \min_{w \in W_{filtered}} \Delta_i(w) \quad (5)$$

Finally, PromptObfus substitutes the privacy-sensitive word at position i with the optimal replacement w^*, iterating this procedure across all masked positions. This sequential filling approach selects each replacement by considering both local contextual constraints and global semantic coherence from prior substitutions, thereby preserving task utility while preserving text semantics.

Experiments Setup

We evaluate the effectiveness of PromptObfus across two critical dimensions, emphasizing its capacity to maintain robust privacy protection while preserving task utility. To demonstrate its practical utility, we apply PromptObfus to three NLP tasks: sentiment analysis, topic classification, and question answering. These tasks represent diverse real-world applications and provide a comprehensive assessment of the method's applicability.

To evaluate PromptObfus's privacy protection capabilities, we simulate adversarial attacks to assess whether sensitive information can be extracted from desensitized prompts. We consider three attack strategies, including two text reconstruction methods and one privacy inference method: Embedding Inference (EI), Mask Token Inference (MTI), and PII Inference. *EI* (Qu et al. 2021) measures the semantic similarity between each word representation and a publicly available word embedding matrix, predicting sensitive content based on the nearest neighbors. *MTI* (Yue et al. 2021) masks tokens in desensitized prompts and assesses the attacker's success in reconstructing the original text. *PII Inference* (Plant, Gkatzia, and Giuffrida 2021) examines textual patterns to deduce private user attributes.

Dataset	Split	Number of Samples
SST-2	Train	67,349
	Validation	872
	Test	1,821
AG News	Train	120,000
	Validation	7,600
	Test	7,600
PersonalPortrait	Test	400

Table 1: Statistics of the datasets.

Baselines

We compare PromptObfus against six state-of-the-art privacy-preserving methods and the original unprotected text. 1) **Random Perturbation**, which randomly substitutes a portion of tokens in the text with arbitrary words. 2) **Presidio** (Microsoft 2025), an automated tool for detecting and redacting sensitive information, including names, locations, and other personally identifiable information. 3) **SANTEXT** (Yue et al. 2021), a differential privacy approach that determines word replacement probabilities based on Euclidean distances in embedding space. 4) **SANTEXT+** (Yue et al. 2021), an improved variant of SANTEXT that incorporates word frequency information to optimize replacement probabilities. 5) **DP Prompt** (Utpala, Hooker, and Chen 2023), a method that employs LLMs to paraphrase original prompts while preserving privacy. 6) **PromptCrypt** (Lin, Hua, and Zhang 2024), which transforms original prompts into emoji sequences using large models.

Evaluation Metrics

Privacy Protection Metrics. We measure the potential leakage of private information to third-party attackers through quantitative evaluation. Two key metrics are adopted to assess privacy protection performance: *TopK Accuracy* and *Success Rate*. *TopK Accuracy* (Zhou et al. 2023) evaluates token-level privacy by computing the proportion of correctly inferred words among the top *k* predictions generated by third-party attackers. *Success Rate* (Plant, Gkatzia, and Giuffrida 2021) measures the exposure risk of personally identifiable information by determining the percentage of successfully extracted PII entities relative to the total identifiable information present.

Task Utility Metrics. To assess PromptObfus's capability in preserving task utility, we measure the model's accuracy when processing desensitized prompts. Our evaluation employs two standard metrics: accuracy and answer quality score. *Accuracy* quantifies the proportion of correct predictions relative to the total number of test instances, applicable to both classification and question answering tasks. *Answer Quality Score* evaluates the overall quality of responses, considering factors including correctness, relevance, completeness, and readability.

Approach	Acc.↑	MTI Top1↓	EI Top1↓	PI Success Rate↓	Avg. Ranking↓
Origin	87.50	31.37	–	–	–
Random	83.75 (4)	17.10 (2)	83.78 (7)	97.50 (9)	5.50
Presidio	83.25 (6)	23.28 (5)	71.53 (6)	**0.00 (1)**	4.50
SANTEXT	61.50 (8)	21.43 (3)	62.10 (5)	41.75 (7)	5.75
SANTEXT+	55.25 (9)	**11.04 (1)**	**49.09 (1)**	34.25 (6)	4.25
DP-Prompt	85.00 (2)	–	–	96.25 (8)	5.00
PromptCrypt	72.00 (7)	–	–	13.50 (5)	6.00
PromptObfus (k=0.1)	**85.25 (1)**	24.68 (7)	61.86 (4)	**0.00 (1)**	3.25
PromptObfus (k=0.2)	84.50 (3)	23.31 (6)	55.82 (3)	**0.00 (1)**	3.25
PromptObfus (k=0.3)	83.75 (4)	22.89 (4)	49.65 (2)	**0.00 (1)**	**2.75**

Table 2: Performance of privacy protection and task utility with detailed rankings on the AG News topic classification task. In the PI Attack, the AG News dataset does not explicitly label privacy attributes. Therefore, the attack assumes that named entities (*e.g.*, person names, locations) represent explicit privacy attributes and targets these for evaluation. The individual rankings are indicated in ( ).

Approach	Acc.↑	Quality Score↑	MTI Top1↓	EI Top1↓	PI(Loc.)↓	PI(Occ.)↓	Avg. Ranking↓
Origin	96.9	3.86	46.43	–	94.75	60.25	–
Random	90.0 (8)	3.34 (6)	**32.67 (1)**	90.00 (6)	81.50 (8)	46.25 (5)	5.67
Presidio	**96.9 (1)**	3.56 (4)	44.16 (5)	96.62 (7)	**0.00 (1)**	55.00 (8)	4.33
SANTEXT	91.0 (6)	3.27 (8)	55.75 (6)	78.56 (4)	**0.00 (1)**	47.00 (6)	5.17
SANTEXT+	91.3 (5)	3.33 (7)	55.75 (6)	**61.62 (1)**	**0.00 (1)**	48.25 (7)	4.50
DP-Prompt	95.0 (3)	3.62 (2)	–	–	89.25 (9)	55.25 (9)	5.75
PromptCrypt	49.5 (9)	2.89 (9)	–	–	16.25 (7)	**11.00 (1)**	6.50
PromptObfus (k=0.1)	96.0 (2)	**3.63 (1)**	42.30 (4)	86.45 (5)	**0.00 (1)**	45.75 (4)	**2.83**
PromptObfus (k=0.2)	93.0 (4)	3.61 (3)	38.81 (3)	77.02 (3)	**0.00 (1)**	37.75 (3)	**2.83**
PromptObfus (k=0.3)	90.5 (7)	3.46 (5)	36.57 (2)	68.10 (2)	**0.00 (1)**	17.25 (2)	3.16

Table 3: Performance of privacy protection and task utility with detailed rankings on the PersonalPortrait text QA task. In the PI Attack, we annotate two types of private entities in the data: Location and Occupation. Among them, Occupation is considered implicit privacy, as it is more likely to be inferred from the context. The individual rankings are indicated in ( ).

Datasets

Our evaluation employs two established benchmark datasets: **SST-2** (Socher et al. 2013) for sentiment analysis and **AG News** (Zhang, Zhao, and LeCun 2015) for topic classification. Since existing QA datasets typically contain anonymized or desensitized content and are therefore unsuitable for privacy evaluation, we develop **PersonalPortrait**, a specialized dataset comprising 400 sensitive psychological counseling dialogues. These patient narratives are generated using GPT-4 and subsequently validated through rigorous manual review by two domain experts to ensure both authenticity and privacy relevance. Complete dataset statistics are presented in Table 1.

Implementation Details

We implement PromptObfus by utilizing three open-source language models: RoBERTa-base (Facebook 2024) as the core desensitization model, BART-large (Facebook 2022) as the task-specific surrogate model for classification tasks, and GPT-Neo-1.3B (EleutherAI 2023) as the general surrogate model for question answering tasks, selected based on dataset size considerations.

To ensure a fair comparison, we maintain a consistent obfuscation ratio across all word-level protection baselines and PromptObfus. As DP Prompt and PromptCrypt operate at the prompt level rather than the word level, they are evaluated solely using PI Attack rather than MTI or EI Attack. All experiments employ the original parameter configurations from their respective publications, with GPT-4o-mini implemented as the remote LLM.

Results and Analysis

Overall Performance

Tables 2 and 3 present the experimental results on the AG News and PersonalPortrait datasets. PromptObfus demonstrates superior performance with an average ranking of 2.75 and 2.83, respectively, surpassing all baseline methods.

Privacy Protection. The PI attack of explicit privacy attains a 0.00% success rate against PromptObfus-generated prompts, confirming complete explicit privacy preservation. Comparative methods (SANTEXT+, DP Prompt, PromptCrypt) exhibit significantly higher vulnerability, as they modify linguistic structures rather than implementing targeted PII protection. In the PI Inference of *Occupation*, PromptObfus achieves the second-lowest attack success rate

Approach	Acc.↑	MTI Top1↓	EI Top1↓
Original Data	87.20	48.86	–
GPT2-base	84.00	42.34	81.80
Roberta-base	84.80	42.66	81.71
BART-base	86.40	42.47	81.73
GPT2-medium	84.53	43.99	81.75
Roberta-large	85.87	42.51	81.71
BART-large	**86.67**	42.94	81.72
Llama-2-7B	84.80	42.47	81.75
ChatGLM3-6B	84.53	42.94	81.72

Table 4: Impact of surrogate model variations on obfuscation effectiveness in sentiment analysis.

Approach (k=0.1)	Acc.↑	MTI Top1↓	EI Top1↓
PromptObfus	**85.25**	24.68	**61.86**
Random masking	84.25	**24.33**	63.98

Table 5: Performance of privacy protection and task utility on the AGNews topic classification task evaluated under random masking and PromptObfus.

at 17.25%, trailing only PromptCrypt (11.00%). Compared against high-accuracy baselines exceeding 90% accuracy, PromptObfus attains a 62.70% decrease in implicit privacy inference attack success rates.

Task Utility Preservation. In the topic classification task (AG News, Table 2), PromptObfus maintains 85.25% classification accuracy at $k = 0.1$, only a 2.57% decrease from the original text. This performance exceeds that of other word-level protection techniques, such as Presidio (83.25%) and SANTEXT+ (55.25%). In the PersonalPortrait dataset (QA task, Table 3), PromptObfus achieves 96.0% accuracy, closely matching the original text's performance (96.9%) with merely 0.93% degradation.

These results collectively indicate that PromptObfus successfully achieves robust privacy protection against remote LLM attacks while preserving the original model's task performance, establishing an optimal privacy-utility tradeoff among all evaluated methods.

Ablation Studies

Impact of Surrogate Model. We investigate the impact of architectures and scales of the surrogate model across three model types: encoder-only (RoBERTa), decoder-only (GPT2), and encoder-decoder (BART) architectures. The evaluation spans three model sizes: base (∼130M parameters, *e.g.*, RoBERTa-base), medium (∼350M parameters, *e.g.*, RoBERTa-large, BART-large, GPT-2-medium), and large (Llama-2-7B, ChatGLM3-6B). Limited by computational resources, we employ full-parameter fine-tuning for small and medium models, while utilizing Low-Rank Adaptation (LoRA) for large models.

The experimental results for sentiment analysis are presented in Table 4. We observe that privacy protection efficacy remains unaffected by either the architecture or scale

Approach (k=0.1)	Acc.↑	MTI Top1↓	EI Top1↓
PromptObfus	**85.25**	**24.68**	**61.86**
Top-1 Selection	84.75	35.06	79.50
Random Selection	83.75	25.14	66.52
<MASK>	83.25	27.55	63.96

Table 6: Performance of privacy protection and task utility on the AGNews topic classification task evaluated under different strategies for selecting the candidate desensitized words.

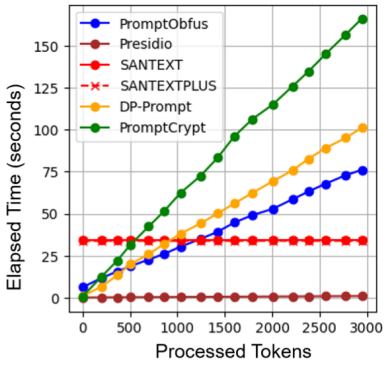

Figure 4: Elapsed time increases linearly with the number of processed tokens across different methods.

of the surrogate model. Medium-scale models demonstrate superior performance compared to their larger counterparts, as the task complexity does not warrant additional model capacity, and LoRA may limit fine-tuning effectiveness. Encoder-decoder architectures achieve optimal performance by effectively integrating the encoder's classification capabilities with the decoder's alignment to remote model requirements.

Impact of Masking Strategy. We examine the effectiveness of different masking strategies in preventing implicit privacy leakage. Our comparison focuses on two approaches: random masking, where tokens are selected uniformly at random, and PromptObfus, a TF-IDF-based method that targets the least frequent tokens. Experimental results on the AGNews dataset reveal that PromptObfus achieves superior performance in both privacy protection and utility preservation compared to random masking, as detailed in Table 5.

Impact of Gradient Filtering. We evaluate four candidate selection strategies for word desensitization: (1) PromptObfus's default gradient-based strategy, which minimizes downstream task impact by selecting tokens with the smallest output gradient magnitudes while excluding original words and their synonyms through additional filtering; (2) top-1 prediction based on model confidence; (3) random selection from candidates; and (4) direct '<MASK>' token insertion as a baseline. We set the number of candidate desensitized words (λ) to 10. The experimental results on the AGNews dataset are shown in Table 6, which demonstrate that PromptObfus's gradient-based approach achieves an optimal balance between privacy preservation and task utility.

Original Text:	I'm a 39 -year-old driver in Toronto , and I often feel like my emotions are all over the place...
Random:	abuser a 39 -year-old driver in Toronto , moha palmery often feel like my emotions are all over shady place...
Presidio:	I'm a <DATE> driver in <GPE>, and I often feel like my emotions are all over the place...
SANTEXT:	jagger rehashed a hardy - year - old driver in women , and obscure often feel like my emotions are all over the place...
SANTEXT+:	jagger rehashed a fidel 15 year 3 old driver in motion , and esoteric seldom feel like my emotions are all putting the however...
DP-Prompt:	I'm a 39 -year-old driver in Toronto , and my emotions can be unpredictable...
PromptCrypt:	39 🚙 🍽️, 😅 → 👺, 😔 → 😞 💔 -> 💔 🚻,...
PromptObfus (k=0.1):	I'm a commercial driver of two and I often feel like my emotions are all over the place...
PromptObfus (k=0.2):	I'm a commercial assistant in LA and I often feel like my emotions flow all over the world...
PromptObfus (k=0.3):	I'm one professional assistant in general and I often feel like my emotions are hovering throughout...

Table 7: A case of desensitized prompts generated by various methods for question answering.

Model	GPT-4o-mini	GLM-4-plus	Meta AI
GPT2	84.53	91.2	**91**
ChatGLM3-6B	84.53	91.0	89
Llama2-7B	84.80	90.8	90
BART	**86.67**	**91.4**	**91**

Table 8: Classification accuracy of local-remote model combinations on the sentiment analysis (SST) task. Columns denote remote models, while rows denote local models.

Time Efficiency Evaluation

We evaluate PromptObfus's computational efficiency on an NVIDIA RTX 3090 GPU with CUDA v12.4. All comparative methods are executed under identical configurations to ensure fair comparison. The results are presented in Figure 4. Our method achieves an optimal balance between computational efficiency and privacy preservation. Notably, the system exhibits a processing rate of 100 tokens in 2.58 seconds, demonstrating practical runtime performance for real-world applications.

Transferability

We further explore the transferability of trained surrogate models across different platform combinations. Experiments evaluate local-remote model pairings from three providers: OpenAI, Meta, and Zhipu. Experimental results, presented in Table 8, indicate that cross-platform model combinations maintain comparable obfuscation effectiveness, showing strong transferability across vendors. For additional validation, we test BART-large, the best-performing independent model from previous experiments, with all three remote models. The results consistently show BART-large's superior performance in every configuration.

Case Study

Table 7 illustrates an example of desensitized prompts generated by various methods for question-answering. The original text contains identifiable sensitive information, including age ('39-year-old'), occupation ('driver'), and location ('Toronto'). PromptObfus successfully replaces explicit private attributes (age, location) with de-identified terms, ensuring robust privacy protection. At $k = 0.2$ and $k = 0.3$, the obfuscation intensity increases, and implicit privacy details, such as occupation ('driver'), are substituted with more ambiguous terms like 'assistant' while preserving semantic coherence and readability.

In contrast, the Random method fails to accurately identify and modify sensitive information, leading to the leakage of all privacy-related terms and a lack of textual coherence. Presidio is limited to handling predefined temporal and geographic patterns, offering insufficient flexibility and failing to protect occupation-related privacy. Meanwhile, SANTEXT and SANTEXT+ introduce excessive noise, rendering the sentences overly disordered and degrading task utility. DP-Prompt results in privacy leakage, while PromptCrypt, though privacy-preserving, employs overly simplistic and abstract symbols, causing significant performance degradation.

Conclusion

In this paper, we propose PromptObfus, a novel method for privacy-preserving prompt desensitization in LLMs based on *anti-adversarial learning*. By replacing sensitive terms with semantically distant but task-consistent alternatives, PromptObfus effectively prevents human interpretation of private content while maintaining output fidelity. Specifically, for each privacy-sensitive word, our method generates multiple candidate substitutions that diverge semantically from the original. A surrogate model then evaluates these candidates to determine their impact on task performance, enabling the selection of desensitized words that least disrupt functional accuracy. Evaluations across three NLP tasks confirm that PromptObfus effectively protects privacy against cloud-based LLM attacks while maintaining utility, outperforming existing baselines in privacy-utility balance.

Acknowledgments

This research is funded by the National Key Research and Development Program of China (Grant No. 2023YFB4503802), the National Natural Science Foundation of China (Grant No. 62032004), and the Natural Science Foundation of Shanghai (Grant No. 25ZR1401175).

References

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2890–2896. Brussels, Belgium: Association for Computational Linguistics.

Chen, Y.; Li, T.; Liu, H.; and Yu, Y. 2023. Hide and Seek (HaS): A Lightweight Framework for Prompt Privacy Protection. arXiv:2309.03057.

Chowdhury, A. R.; Glukhov, D.; Anshumaan, D.; Chalasani, P.; Papernot, N.; Jha, S.; and Bellare, M. 2025. Pr$\epsilon\epsilon$mpt: Sanitizing Sensitive Prompts for LLMs. arXiv:2504.05147.

Duan, H.; Dziedzic, A.; Yaghini, M.; Papernot, N.; and Boenisch, F. 2024. On the Privacy Risk of In-context Learning. arXiv:2411.10512.

EleutherAI. 2023. GPT-Neo 1.3B. Https://huggingface.co/EleutherAI/gpt-neo-1.3B.

Explosion. 2024. spaCy's named entity recognition model. Https://spacy.io/models/en/#en_core_web_trf.

Facebook. 2022. BART (large-sized model). Https://huggingface.co/facebook/bart-large.

Facebook. 2024. RoBERTa base model. Https://huggingface.co/FacebookAI/roberta-base.

Frikha, A.; Razi, M. R. A.; Nakka, K. K.; Mendes, R.; Jiang, X.; and Zhou, X. 2025. PrivacyScalpel: Enhancing LLM Privacy via Interpretable Feature Intervention with Sparse Autoencoders. arXiv:2503.11232.

Gao, F.; Zhou, R.; Wang, T.; Shen, C.; and Yang, J. 2025. Data-adaptive Differentially Private Prompt Synthesis for In-Context Learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Gentry, C. 2009. *A fully homomorphic encryption scheme*. Ph.D. thesis, Stanford University, Stanford, CA, USA. AAI3382729.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, 2672–2680. Cambridge, MA, USA: MIT Press.

Hao, M.; Li, H.; Chen, H.; Xing, P.; Xu, G.; and Zhang, T. 2022. Iron: Private Inference on Transformers. In *Advances in Neural Information Processing Systems*, volume 35, 15718–15731.

Hu, J.; Wang, D.; Wang, Z.; Pang, X.; Xu, H.; Ren, J.; and Ren, K. 2024. Federated Large Language Model: Solutions, Challenges and Future Directions. *IEEE Wireless Communications*, 1–8.

Lin, G.; Hua, W.; and Zhang, Y. 2024. EmojiCrypt: Prompt Encryption for Secure Communication with Large Language Models. arXiv:2402.05868.

Liu, X.-Y.; Zhu, R.; Zha, D.; Gao, J.; Zhong, S.; White, M.; and Qiu, M. 2025. Differentially Private Low-Rank Adaptation of Large Language Model Using Federated Learning. *ACM Trans. Manage. Inf. Syst.*, 16(2).

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and Arcas, B. A. y. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, 1273–1282. PMLR.

Microsoft. 2025. Presidio: Data Protection and De-identification SDK. Https://microsoft.github.io/presidio/.

Miranda, M.; Ruzzetti, E. S.; Santilli, A.; Zanzotto, F. M.; Bratières, S.; and Rodolà, E. 2025. Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions. *Trans. Mach. Learn. Res.*, 2025.

Mireshghallah, N.; Kim, H.; Zhou, X.; Tsvetkov, Y.; Sap, M.; Shokri, R.; and Choi, Y. 2024. Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Plant, R.; Gkatzia, D.; and Giuffrida, V. 2021. CAPE: Context-Aware Private Embeddings for Private Language Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7970–7978. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Pryzant, R.; Iter, D.; Li, J.; Lee, Y.; Zhu, C.; and Zeng, M. 2023. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7957–7968. Singapore: Association for Computational Linguistics.

Qu, C.; Kong, W.; Yang, L.; Zhang, M.; Bendersky, M.; and Najork, M. 2021. Natural Language Understanding with Privacy-Preserving BERT. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, 1488–1497. New York, NY, USA: Association for Computing Machinery.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.

Sparck Jones, K. 1988. *A statistical interpretation of term specificity and its application in retrieval*, 132–142. GBR: Taylor Graham Publishing. ISBN 0947568212.

Sun, X.; Liu, G.; He, Z.; Li, H.; and Li, X. 2024. DePrompt: Desensitization and Evaluation of Personal Identifiable Information in Large Language Model Prompts. arXiv:2408.08930.

Utpala, S.; Hooker, S.; and Chen, P.-Y. 2023. Locally Differentially Private Document Generation Using Zero Shot Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8442–8457. Singapore: Association for Computational Linguistics.

Vats, A.; Liu, Z.; Su, P.; Paul, D.; Ma, Y.; Pang, Y.; Ahmed, Z.; and Kalinli, O. 2024. Recovering from Privacy-Preserving Masking with Large Language Models. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10771–10775.

Wang, Z.; Wang, W.; Chen, Q.; Wang, Q.; and Nguyen, A. 2024. Generating Valid and Natural Adversarial Examples with Large Language Models. In *27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024, Tianjin, China, May 8-10, 2024*, 1716–1721. IEEE.

Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q. V.; Zhou, D.; and Chen, X. 2024. Large Language Models as Optimizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Yao, A. C. 1982. Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science*, 160–164.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, 4(2).

Yue, X.; Du, M.; Wang, T.; Li, Y.; Sun, H.; and Chow, S. S. M. 2021. Differential Privacy for Text Analytics via Natural Text Sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3853–3866. Online: Association for Computational Linguistics.

Zhang, X.; Pang, Y.; Kang, Y.; Chen, W.; Fan, L.; Jin, H.; and Yang, Q. 2025. No free lunch theorem for privacy-preserving LLM inference. *Artif. Intell.*, 341: 104293.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, 649–657. Cambridge, MA, USA: MIT Press.

Zhou, X.; Lu, Y.; Ma, R.; Gui, T.; Wang, Y.; Ding, Y.; Zhang, Y.; Zhang, Q.; and Huang, X. 2023. TextObfuscator: Making Pre-trained Language Model a Privacy Protector via Obfuscating Word Representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, 5459–5473. Toronto, Canada: Association for Computational Linguistics.

Zhou, Y.; Muresanu, A. I.; Han, Z.; Paster, K.; Pitis, S.; Chan, H.; and Ba, J. 2022. Large Language Models Are Human-Level Prompt Engineers. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.