Anti-adversarial Learning: Desensitizing Prompts for Large Language Models

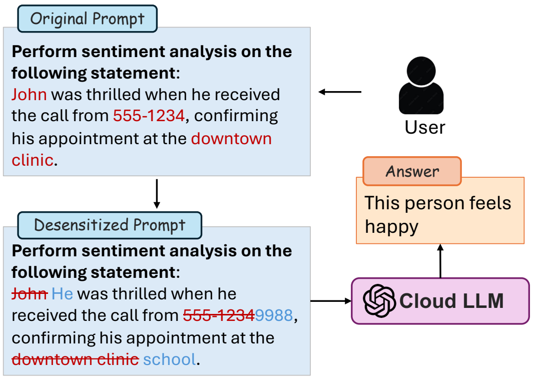
Anonymous ACL submission

Abstract

With the widespread use of LLMs, preserving privacy in user prompts has become crucial, as prompts risk exposing privacy and sensitive data to the cloud LLMs. Conventional techniques like homomorphic encryption, secure multi-party computation, and federated learning face challenges due to heavy computational overhead and user participation demands, limiting their applicability in LLM scenarios. In this paper, we propose PromptObfus, a novel method for desensitizing LLM prompts. The core idea of PromptObfus is "anti-adversarial" learning, which perturbs privacy words in the prompt to obscure sensitive information while retaining the stability of model 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. Our code is publicly available at https://anonymous.4open.science/r/PromptObfus-BF36/.

1 Introduction

The widespread adoption of large language models (LLMs) such as ChatGPT in various NLP tasks (Hong et al., 2024; Carlini et al., 2019) 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.



Perform sentiment analysis on the following statement: John was thrilled when he received the call from 555-1234, confirming his appointment at the downtown clinic.

This person feels happy

Perform sentiment analysis on the following statement: ~~John~~ He was thrilled when he received the call from ~~555-1234~~9988, confirming his appointment at the ~~downtown clinic~~ school.

Figure 1: Illustration of prompt desensitization.

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

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., 2024), 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., 2024). Furthermore, 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 aims to obscure sensitive words in prompts while preserving the integrity of model predictions. 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.

2 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 et al., 2024) and noise-based obfuscation (Zhou et al., 2023; Gao et al., 2024); 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 et al., 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., 2023).

In contrast to existing approaches, we propose

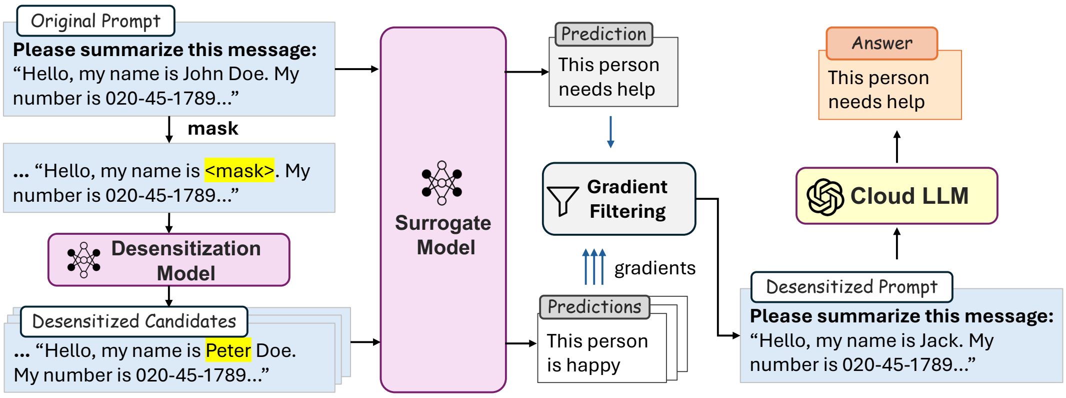

Original Prompt
Please summarize this message:
"Hello, my name is John Doe. My number is 020-45-1789..."
mask
... "Hello, my name is <mask>. My number is 020-45-1789..."
Desensitization Model
Desensitized Candidates
... "Hello, my name is Peter Doe. My number is 020-45-1789..."
Surrogate Model
Prediction
This person needs help
Gradient Filtering
gradients
Predictions
This person is happy
Answer
This person needs help
Cloud LLM
Desensitized Prompt
Please summarize this message:
"Hello, my name is Jack. My number is 020-45-1789..."

Figure 2: Overview of PromptObfus.

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.

3 Approach

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.

3.1 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 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:

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

Here, $M(x|\lambda, k)$ denotes a desensitization mapping function, where 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.

3.2 Overview

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

3.3 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 named entity recognition (NER) model[1] to detect explicit privacy attributes like person names, locations, and organizations. All identified privacy-sensitive words are uniformly

[1] https://spacy.io/models/en/#en_core_web_trf

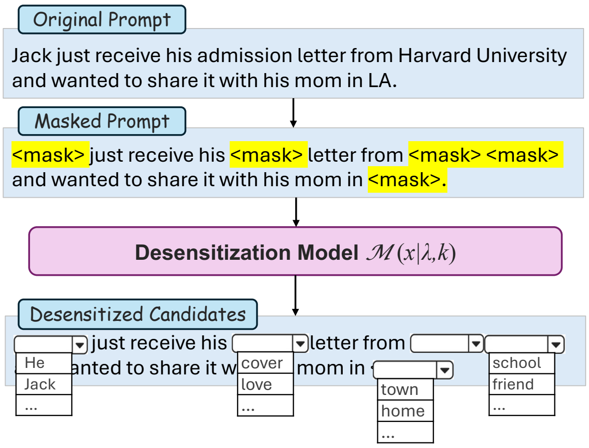

Original Prompt
Jack just receive his admission letter from Harvard University and wanted to share it with his mom in LA.
Masked Prompt
<mask> just receive his <mask> letter from <mask> <mask> and wanted to share it with his mom in <mask>.
Desensitization Model $\mathcal{M}(x|\lambda,k)$
Desensitized Candidates
just receive his
letter from
anted to share it w
mom in
He
Jack
...
cover
love
...
town
home
...
school
friend
...

Figure 3: Illustration of predicting candidate desensitive words.

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. For each candidate word w_i, we calculate its Euclidean distance to the original words x_{original} using word embeddings:

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

where $\vec{x_{\text{original}}}$ and $\vec{w_i}$ represent the word vectors of the original and desensitized words, respectively, and $\| \cdot \|$ denotes the Euclidean norm.

A distance threshold θ_{dist} filters the desensitized word set. Words with Euclidean distance $d(x_{\text{original}}, w_i) \leq \theta_{\text{dist}}$ are removed for semantic similarity. The filtered set is defined as:

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

where W_{filtered} denotes the refined candidate set after exclusion.

3.4 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. 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.

3.5 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:

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

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:

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

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.

4 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 et al., 2021) examines textual patterns to deduce private user attributes.

4.1 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**[2], 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 et al., 2023), a method that employs LLMs to paraphrase original prompts while preserving privacy. 6) **PromptCrypt** (Lin et al., 2024), which transforms original prompts into emoji sequences using large models.

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.

4.2 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 et al., 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. For automated assessment, we employ GPT-4o-mini as an evaluator, with the complete scoring rubric provided in Appendix A.2.

4.3 Datasets

Our evaluation employs two established benchmark datasets: **SST-2** (Socher et al., 2013) for sentiment analysis and **AG News** (Zhang et al., 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

[2] https://microsoft.github.io/presidio/

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 sentiment analysis 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 ( ).

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, while the detailed construction process is documented in Appendix A.1.

4.4 Implementation Details

We implement PromptObfus by utilizing three open-source language models: RoBERTa-base[3] as the core desensitization model, BART-large[4] as the task-specific surrogate model for classification tasks, and GPT-Neo-1.3B[5] 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. Further details on hyperparameter configurations are provided in Appendix A.3.

5 Results and Analysis

5.1 Overall Performance

Table 2 presents the experimental results on the AG News dataset (further details for other datasets are provided in Appendix A.4). PromptObfus ($k = 0.3$) demonstrates superior performance with an average ranking of 2.75, surpassing all baseline methods.

Privacy Protection. The PI attack attains a 0.00% success rate against PromptObfus-generated prompts, confirming complete privacy preservation. Comparative methods (SANTEXT+, DP Prompt, PromptCrypt) exhibit significantly higher vulnerability, as they modify linguistic structures rather than implementing targeted PII protection. For EI Attack, PromptObfus ($k = 0.3$) achieves a 49.65% success rate, outperforming all baselines except SANTEXT+.

Task Utility Preservation. PromptObfus maintains 85.25% classification accuracy at $k = 0.1$, representing 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%).

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.

5.2 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 ($\sim$130M parameters, e.g., RoBERTa-base), medium ($\sim$350M parameters, e.g., RoBERTa-large, BART-large, GPT-2-

[3] https://huggingface.co/FacebookAI/roberta-base

[4] https://huggingface.co/facebook/bart-large

[5] https://huggingface.co/EleutherAI/gpt-neo-1.3B

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 3: 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 4: Performance of privacy protection and task utility on the AGNews topic classification task evaluated under random masking and PromptObfus.

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

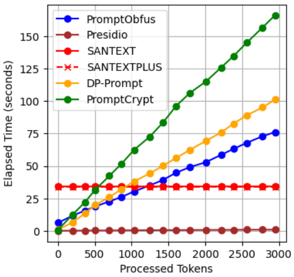


Figure 4: Elapsed time increases linearly with the number of processed tokens across different methods.

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 3, with corresponding question answering results provided in Appendix A.5. We observe that privacy protection efficacy remains unaffected by either the architecture or scale 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 4.

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 smallest output gradient magnitudes; (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 5, demonstrating that PromptObfus's gradient-based approach achieves an optimal balance between privacy preservation and task utility.

Additionally, a detailed examination of hyperparameters k and λ is presented in Appendix A.6.

5.3 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.

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 6: 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 7: Classification accuracy of local-remote model combinations on the sentiment analysis (SST) task. Columns denote remote models, while rows denote local models.

5.4 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 7, 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.

5.5 Case Study

Table 6 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, despite protecting privacy, employs overly simplistic and abstract symbols, causing significant performance degradation.

6 Conclusion

In this paper, we introduces PromptObfus, a novel method for privacy-preserving prompt desensitization in LLMs. Its core idea is *anti-adversarial learning*, which simultaneously preserves model output fidelity while preventing human interpretation of sensitive content. PromptObfus achieves this by replacing sensitive words in user prompts with semantically distant yet task-consistent alternatives, minimizing impact on task utility. Evaluations across three NLP tasks demonstrate PromptObfus's effectiveness in safeguarding privacy against cloud-based LLM attacks while maintaining original task utility levels. The results establish PromptObfus's superior privacy-utility balance compared to existing baseline methods.

Limitations

We have identified two limitations of PromptObfus:

Partial coverage of implicit privacy attributes: Our approach identifies potential privacy-related terms through TF-IDF scoring, which effectively reduces but does not fully eliminate privacy leakage. This limitation stems from incomplete detection of all sensitive attributes. Future improvements should develop more sophisticated privacy attribute detection mechanisms.

General model constraints: The reliance on general-purpose models, necessitated by limited annotated QA datasets, results in suboptimal performance compared to task-specific models. Investigating few-shot learning approaches (Brown et al., 2020) could address this data scarcity challenge in future work.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 267–284, USA. USENIX Association.

Yu Chen, Tingxin Li, Huiming Liu, and Yang Yu. 2023. Hide and seek (has): A lightweight framework for prompt privacy protection. *Preprint*, arXiv:2309.03057.

Amrita Roy Chowdhury, David Glukhov, Divyam Anshumaan, Prasad Chalasani, Nicolas Papernot, Somesh Jha, and Mihir Bellare. 2025. Pr$\epsilon\epsilon$mpt: Sanitizing sensitive prompts for llms. *Preprint*, arXiv:2504.05147.

Ahmed Frikha, Muhammad Reza Ar Razi, Krishna Kanth Nakka, Ricardo Mendes, Xue Jiang, and Xuebing Zhou. 2025. Privacyscalpel: Enhancing llm privacy via interpretable feature intervention with sparse autoencoders. *Preprint*, arXiv:2503.11232.

Fengyu Gao, Ruida Zhou, Tianhao Wang, Cong Shen, and Jing Yang. 2024. Data-adaptive differentially private prompt synthesis for in-context learning. *Preprint*, arXiv:2410.12085.

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2672–2680, Cambridge, MA, USA. MIT Press.

Meng Hao, Hongwei Li, Hanxiao Chen, Pengzhi Xing, Guowen Xu, and Tianwei Zhang. 2022. Iron: Private inference on transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 15718–15731. Curran Associates, Inc.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng LI, Bo Li, and Zhangyang Wang. 2024. DP-OPT: Make large language model your privacy-preserving prompt engineer. In *The Twelfth International Conference on Learning Representations*.

Jiahui Hu, Dan Wang, Zhibo Wang, Xiaoyi Pang, Huiyu Xu, Ju Ren, and Kui Ren. 2024. Federated large language model: Solutions, challenges and future directions. *IEEE Wireless Communications*, pages 1–8.

Guo Lin, Wenyue Hua, and Yongfeng Zhang. 2024. Emojicrypt: Prompt encryption for secure communication with large language models. *Preprint*, arXiv:2402.05868.

Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. 2025. Differentially private low-rank adaptation of large language model using federated learning. *ACM Trans. Manage. Inf. Syst.*, 16(2).

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 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*, pages 1273–1282. PMLR.

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, Fabio Massimo Zanzotto, Sébastien Bratières, and Emanuele Rodolà. 2025. Preserving privacy in large language models: A survey on current threats and solutions. *Preprint*, arXiv:2408.05212.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *Preprint*, arXiv:2310.17884.

Richard Plant, Dimitra Gkatzia, and Valerio Giuffrida. 2021. CAPE: Context-aware private embeddings for private language learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7970–7978, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural language understanding with privacy-preserving bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 1488–1497, New York, NY, USA. Association for Computing Machinery.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

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

Robin Staab, Mark Vero, Mislav Balunovi'c, and Martin T. Vechev. 2023. Beyond memorization: Violating privacy via inference with large language models. *ArXiv*, abs/2310.07298.

Xiongtao Sun, Gan Liu, Zhipeng He, Hui Li, and Xiaoguang Li. 2024. Deprompt: Desensitization and evaluation of personal identifiable information in large language model prompts. *Preprint*, arXiv:2408.08930.

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

Arpita Vats, Zhe Liu, Peng Su, Debjyoti Paul, Yingyi Ma, Yutong Pang, Zeeshan Ahmed, and Ozlem Kalinli. 2024. Recovering from privacy-preserving masking with large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10771–10775.

Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2023. Generating valid and natural adversarial examples with large language models. *Preprint*, arXiv:2311.11861.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. *Preprint*, arXiv:2309.03409.

Andrew C. Yao. 1982. Protocols for secure computations. In *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pages 160–164.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Xiaojin Zhang, Yulin Fei, Yan Kang, Wei Chen, Lixin Fan, Hai Jin, and Qiang Yang. 2024. No free lunch theorem for privacy-preserving llm inference. *Preprint*, arXiv:2405.20681.

Xin Zhou, Yi Lu, Ruotian Ma, Tao Gui, Yuran Wang, Yong Ding, Yibo Zhang, Qi Zhang, and Xuanjing Huang. 2023. TextObfuscator: Making pre-trained language model a privacy protector via obfuscating word representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5459–5473, Toronto, Canada. Association for Computational Linguistics.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

A Appendix

A.1 PersonalPortrait Construction

Inspired by the D4 dataset (Yao et al., 2022) and the PersonalReddit dataset (Staab et al., 2023), which synthesize text from personal profiles, we develop authentic patient profiles incorporating demographic characteristics (gender, occupation, location) and psychiatric conditions to simulate clinical counseling dialogues. The QA task focuses on mental health diagnosis, where models must identify specific disorders (e.g., depression, anxiety) through symptom analysis (e.g., insomnia, persistent sadness, anxious thoughts) present in the counseling transcripts.

The dataset synthesis involves three key stages:

1. Profile generation: Diverse patient profiles are created by randomly sampling demographic attributes (age, gender, location, occupation) and psychiatric conditions from predefined clinical categories.

2. Text generation: GPT-4 synthesizes clinically coherent self-reports based on generated profiles, capturing authentic symptom presentations and personal histories (see Appendix A.2 for generation templates).

3. Clinical validation: Generated texts undergo expert review to verify diagnostic accuracy, symptom plausibility, and therapeutic appropriateness.

Each entry in the PersonalPortrait dataset contains:

- Clinical Narrative: A first-person account simulating psychological counseling sessions, documenting symptom presentation, behavioral patterns, and psychosocial experiences related to mental health conditions.

- Age: Integer values uniformly distributed between 18 and 65 years.

- Location: Randomly sampled from 20 predefined English-speaking geographical regions.

- Gender: Binary classification (male/female) with equal probability.

- Occupation: Profession randomly selected from a standardized list of 20 common occupations.

- Clinical Diagnosis: Mental health disorder randomly assigned from 10 prevalent psychiatric conditions following DSM-5 classification criteria.

A.2 Prompt Templates

We present the core prompt templates utilized in our study, where *TEXT* indicates the desensitized prompt, *ORIGINAL TEXT* corresponds to the unprotected original prompt, and *ANSWER* signifies the LLM-generated response. Attributes such as *AGE* and *GENDER* are randomly sampled from predefined categories.

Prompt for Generating Samples for PersonalPortrait

```
System Prompt:
You are a patient and are seeing a
psychiatrist.

User Prompt:
Describe to a psychiatrist in the
typical voice of a <<<AGE>>>-year-old
<<<GENDER>>> <<<OCCUPATION>>>
with symphons of <<<DISORDER>>>
in <<<LOCATION>>>. The self-report
MUST reflect the patient's LOCATION and
OCCUPATION.100 words or less.
```

Prompt for Sentiment Analysis Task

```
System Prompt:
Classify the sentiment of each sentence
in <<<SENTENCE>>> as 'Positive'
or 'Negative'. Give the sentiment
classifications without any other preamble
text.

###EXAMPLE SENTENCE
Highly recommend this company for travel
plans involving rail.
EXAMPLE OUTPUT
Positive###

User Prompt:
<<<TEXT>>>
```

Prompt for Topic Classification Task

```
System Prompt:
Classify the topic of each news article
in <<<NEWS>>> as 'World', 'Sports',
'Business', or 'Sci/Tech'. Give the topic
classification without any other preamble
text.

###EXAMPLE NEWS
Dollar Briefly Hits 4-Wk Low Vs Euro LONDON
(Reuters) - The dollar dipped to a four-week
low against the euro on Monday before rising
slightly on profit-taking, but steep oil
prices and weak U.S. data continued to fan
worries about the health of the world's
largest economy.
EXAMPLE OUTPUT
Business###

User Prompt:
<<<TEXT>>>
```

Prompt for Question Answering Task

```
System Prompt:
You are an experienced psychologist. Your task is to listen to the patient's self-description and a most likely mental health diagnosis based on their symptoms.

User Prompt:
<<<TEXT>>>
Diagnosis:
```

Prompt for Scoring Answer Quality

```
System Prompt:
You are an experienced psychologist with expertise in mental health diagnoses. Your task is to assess the quality of a diagnosis based on the patient's self-description. You will evaluate the diagnosis on four dimensions, each on a scale from 1 to 5, where 1 indicates poor performance and 5 indicates excellent performance. After evaluating all four dimensions, please also provide an overall score (1 to 5) that reflects your holistic judgment of the diagnosis quality.Please consider the following criteria:
Correctness - How clinically accurate is the diagnosis in light of the symptoms described?
Relevance - How well does the diagnosis relate to and reflect the patient's specific symptoms?
Completeness - Does the diagnosis address all key aspects of the patient's self-report without overlooking significant information?
Readability - Is the diagnosis clearly written, well-structured, and understandable?

User Prompt:
The patient has provided the following self-description. The diagnosis made by the doctor is as follows: TEXT:<<<ORIGINAL TEXT>>> Diagnosis:<<<ANSWER>>> Please provide a score from 1 to 5 for each of the following dimensions: Correctness, Relevance, Completeness, Readability. Then, based on the four criteria above, provide an Overall Score (1 to 5) that reflects your general assessment of the diagnosis.
```

A.3 Hyperparameter Setting

The model training configurations are specified in Tables 8 and 9. For Llama-2-7B and ChatGLM3-6B, we implement Low-Rank Adaptation (LoRA), whereas other models undergo standard full-parameter fine-tuning. Optimization is performed using Adam with standard parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. All model usage strictly adheres to respective licensing agreements.

The experiments are conducted on an Ubuntu 23.10 server with vCUDA 12.4, utilizing an Nvidia GeForce RTX 3090 GPU.

Dataset	Model	lr	bs	epoch
SST-2	Roberta-base	2e-5	32	4
	Roberta-large	3e-5	32	4
	BART-base	2e-5	32	4
	BART-large	3e-5	32	4
	GPT2-base	3e-5	32	4
	GPT2-medium	3e-5	32	4
	Llama-2-7B	2e-4	16	2
	ChatGLM3-6B	2e-4	16	2
AG News	BART-large	3e-5	32	5

Table 8: Hyperparameters setting for model training.

Dataset	Model	alpha	dropout	r
SST-2	Llama-2-7B	16	0.1	64
	ChatGLM3-6B	16	0.1	64

Table 9: LoRA hyperparameters setting for model training.

A.4 Results on Other Datasets

Tables 10 and 11 present the results on the sentiment analysis and question answering tasks, respectively. We can observe that PromptObfus consistently achieves an optimal trade-off between privacy protection and task utility, exhibiting superior performance compared to baseline methods. This performance advantage aligns with the observations from the sentiment analysis task.

Privacy Protection. PromptObfus exhibits superior privacy preservation across both evaluation tasks. For instance, the question answering assessment examines two privacy attributes: *Location* (explicitly stated information) and *Occupation* (implicitly derived sensitive data). In the PI Inference of *Location*, PromptObfus achieves an attack success rate of 0.00%, indicating complete privacy protection. In the PI Inference of *Occupation*, PromptObfus achieves the second-lowest attack success rate 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 sentiment analysis task, PromptObfus maintains strong utility preservation, achieving 86.67% accuracy at $k = 0.1$, comparable to baseline methods (87.20%)

Approach	**Acc.↑**	**MTI Top1↓**	**EI Top1↓**	**PI Success Rate↓**	**Avg. Ranking↓**
Origin	87.20	48.86	–	–	–
Random	69.87 (7)	35.91 (3)	90.47 (7)	83.47 (8)	6.25
Presidio	84.80 (5)	44.63 (7)	90.45 (6)	**0.00 (1)**	4.75
SANTEXT	49.25 (9)	**20.15 (1)**	73.67 (2)	92.53 (9)	5.25
SANTEXT+	58.93 (8)	23.40 (2)	76.93 (4)	75.47 (7)	5.25
DP-Prompt	86.30 (4)	–	–	72.53 (6)	5.00
PromptCrypt	**89.86 (1)**	–	–	54.67 (5)	**3.00**
PromptObfus (k=0.1)	86.67 (2)	42.94 (6)	81.72 (5)	**0.00 (1)**	3.50
PromptObfus (k=0.2)	86.40 (3)	41.48 (5)	74.67 (3)	**0.00 (1)**	**3.00**
PromptObfus (k=0.3)	83.20 (6)	39.68 (4)	**67.15 (1)**	**0.00 (1)**	**3.00**

Table 10: Performance of privacy protection and task utility with detailed rankings on the SST-2 sentiment analysis task. 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 11: Performance of privacy protection and task utility with detailed rankings on the PersonalPortrait text QA task. The individual rankings are indicated in ( ).

with only 0.61% performance degradation. While PromptCrypt (89.86%) shows marginally better results on this simpler task with limited label space, its emoji-based encryption proves particularly suited for such low-complexity scenarios.

In the question answering task, PromptObfus achieves 96.0% accuracy, closely matching the original text's performance (96.9%) with merely 0.93% degradation, ranking second only to Presidio. This performance can be attributed to the task's primary dependence on contextual emotional inference rather than explicit PII extraction. Notably, PromptObfus obtains the highest answer quality score (3.63), demonstrating superior response fluency, completeness, and accuracy compared to alternative methods.

PromptCrypt shows limited effectiveness in preserving QA task utility. While its encryption-based approach successfully disrupts contextual structures to enhance implicit privacy protection, the consequent loss of semantic information significantly impairs its ability to handle tasks requiring nuanced text analysis.

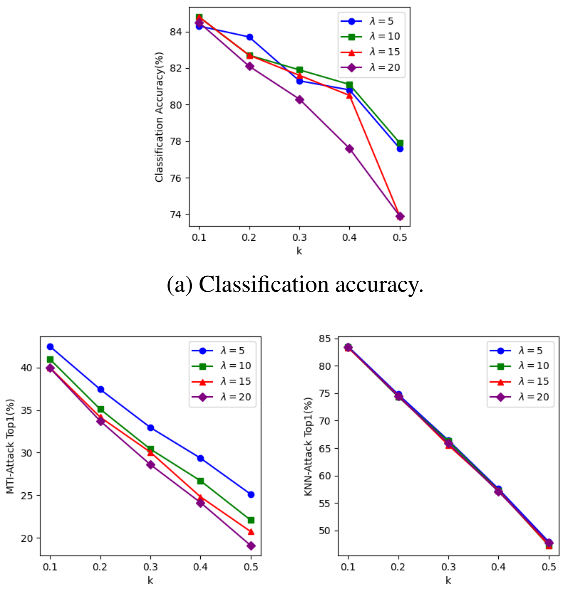
A.5 Impact of Surrogate Model on Other Tasks

Table 12 summarizes the results examining surrogate model effects on question answering performance. As privacy protection effectiveness is previously established to be invariant to surrogate model choice in sentiment analysis, the current evaluation specifically assesses task utility preservation. The investigation utilizes general-purpose surrogate models spanning three architectures of comparable scale (RoBERTa-large, BART-large, and GPT2-medium) alongside a progressively scaled GPT series (GPT2-base, GPT2-medium, and GPT-Neo-1.3B).

The results indicate GPT-Neo-1.3B delivers optimal performance, achieving 96.0% question answering accuracy and the maximal answer quality score. Architectural comparisons reveal GPT2's superior performance over other medium-scale models, confirming the efficacy of decoder-only architectures for generative language tasks. Scaling analysis demonstrates monotonic improvement in question answering accuracy with increasing model size, attributable to larger models' enhanced pre-

Model	Accuracy	Utility Score
GPT2-base	93.3	3.55
GPT2-medium	93.8	3.57
GPTNeo-1.3B	**96.0**	**3.63**
RoBERTa-large	93.0	3.53
BART-large	92.8	3.55

Table 12: Influence of surrogate model variations on obfuscation effectiveness in question answering.



(a) Classification accuracy.

(b) MTI attack.

(c) EI attack.

Figure 5: Impact of hyperparameters k and λ.

trained knowledge representation and superior task execution capacity, particularly beneficial for complex textual question answering scenarios.

A.6 Impact of Hyperparameters

We perform ablation studies on the hyperparameters k and λ, using BART-large as the surrogate model on the SST dataset. The parameter k is varied from 0.1 to 0.5 in increments of 0.1, while λ ranges from 5 to 20 in increments of 5. The results are illustrated in Figure 5.

Regarding privacy protection performance, the Attack Top1 metric decreases monotonically with increasing k, demonstrating improved privacy preservation at higher obfuscation levels. For MTI Attack, larger λ values lead to reduced Top1 scores, with the most substantial enhancement occurring between $\lambda = 5$ and $\lambda = 10$. This improvement stems from more diverse contextual information generating varied MTI predictions. The EI Attack Top1 depends exclusively on k, as this attack analyzes perturbed words independently of their contextual surroundings.

Concerning task utility preservation, classification accuracy exhibits a gradual decline as k increases, with the most pronounced performance degradation observed between $k = 0.4$ and $k = 0.5$. When k exceeds 0.3, the system becomes sensitive to λ variations, where higher values adversely affect performance due to excessive word substitutions compromising semantic integrity and contextual coherence.

Our analysis reveals a fundamental trade-off between privacy protection and task utility with respect to parameters k and λ. While increasing either parameter improves privacy preservation, this comes at the expense of reduced performance. The optimal operating regime occurs when $k \leq 0.4$ and $\lambda \in [10, 20)$, achieving an effective balance between these competing objectives. Based on these findings, we establish $\lambda = 10$ as the default configuration.