Semantic Encryption: Secure and Effective Interaction with Cloud-based Large Language Models via Semantic Transformation

Dong Chen[1,2,3],Tong Yang[1], Feipeng Zhai[1], Pengpeng Ouyang[1], Qidong Liu[1,2,3], Yafei Li[1,2,3], Chong Fu[4], Mingliang Xu[1,2,3,*],

[1]The School of Computer and Artificial Intelligence of Zhengzhou University
[2]Engineering Research Center of Intelligent Swarm Systems, Ministry of Education
[3]National Supercomputing Center In Zhengzhou
[4]College of Computer Science and Technology of Zhejiang University
chendongai@zzu.edu.cn, yangtong@gs.zzu.edu.cn, zhaifeipeng@stu.zzu.edu.cn, oy100253@gs.zzu.edu.cn, ieqdliu@zzu.edu.cn, ieyfli@zzu.edu.cn, fuchong@zju.edu.cn, iexumingliang@zzu.edu.cn

Abstract

The increasing adoption of Cloud-based Large Language Models (CLLMs) has raised significant concerns regarding data privacy during user interactions. While existing approaches primarily focus on encrypting sensitive information, they often overlook the logical structure of user inputs. This oversight can lead to reduced data utility and degraded performance of CLLMs. To address these limitations and enable secure yet effective interactions, we propose Semantic Encryption (SE)—a plug-and-play framework designed to preserve both privacy and utility. SE consists of two key components: Semantic Encoding and Semantic Decoding. In the encoding phase, a lightweight local model transforms the original user input into an alternative semantic context that maintains the original intent and logical structure while obfuscating sensitive information. This transformed input is then processed by the CLLM, which generates a response based on the transformed semantic context. To maintain a seamless user experience, the decoding phase will reconstruct the CLLM's response back into the original semantic context by referencing the locally stored user input. Extensive experimental evaluations demonstrate that SE effectively protects data privacy without compromising data utility or user experience, offering a practical solution for secure interaction with CLLMs. Particularly, the proposed SE demonstrates a significant improvement over the state-of-the-art InferDPT, surpassing it across various evaluated metrics and datasets.

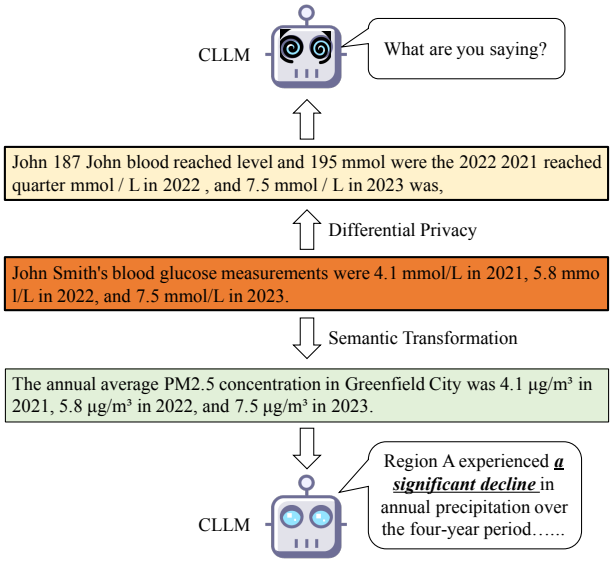

Figure 1: A comparative case study of traditional encryption methods and the proposed semantic transformation in the context of interactions with CLLMs. It is important to emphasize that the examples illustrated in the figure are entirely synthetic and do not contain any personal or sensitive information.

Introduction

Cloud-based Large Language Models (CLLMs), which offer services such as data analysis through Application Programming Interfaces (APIs), are increasingly integrated into everyday life. However, transmitting data to the cloud via APIs for processing and analysis by CLLMs has raised significant concerns regarding data privacy. In particular, service providers may collect user data for model training purposes, further amplifying the risk of privacy leakages (Wang et al. 2024; Wu et al. 2024; Yang et al. 2023).

An increasing body of research has focused on protecting user data privacy during interactions with CLLMs, with particular emphasis on encryption-based techniques (Yan et al. 2024; Yao et al. 2024; Feretzakis et al. 2024). While encryption-based techniques such as differential privacy (Hoory et al. 2021; Du and Mi 2021) provide strong privacy guarantees by introducing randomness, they often come at the cost of reduced data utility, thereby impairing the CLLMs' ability to interpret and analyze user inputs effectively. As illustrated in Figure 1, the orange box presents a patient's blood glucose test record. In the upper section of the figure, the input is encrypted with a differential privacy mechanism. Although the differential privacy method effectively safeguards data privacy, it substantially distorts the original intent and logical structure of the input, resulting in a complete loss of utility and hindering the CLLM's

*Mingliang Xu is the corresponding author.

ability to interpret user input.

Similar to the common practice of safeguarding privacy on public platforms by obscuring only personally identifiable information—such as blurring faces in photographs—it may be unnecessary to encrypt the entirety of user input to ensure data privacy (Chen et al. 2023a; Dong et al. 2024), thereby helping to preserve the input's logical structure and the user's intent. Motivated by this observation, we introduce semantic transformation that analyzes the original user input and transforms it into a logically consistent but semantically different representation. As illustrated in the light green box of Figure 1, semantic transformation converts the patient's blood glucose test record into a context representing annual average PM2.5 concentration record. This transformation can preserve the data utility for tasks such as trend analysis, while fully obscuring all the patient's information, thereby protecting data privacy.

Based on the idea of semantic transformation, we propose Semantic Encryption (SE) that comprises a Semantic Encoder and a Semantic Decoder. More specifically, the Semantic Encoder transforms the original input into an alternative semantic context while preserving its underlying logical structure. Correspondingly, the Semantic Decoder reconstructs the CLLMs' response based on the original input, restoring it to the original semantic context. To improve the deployability of the SE across heterogeneous devices, efficient and lightweight local models are utilized for both the encoder and decoder. Furthermore, we propose Semantic Distillation, a technique that enables local models to effectively learn and replicate the semantic encoding and decoding capabilities of CLLMs. Since all user interactions with the CLLMs are preserved within their original semantic context, the operation of the SE remains virtually imperceptible to users.

This paper proposes a method for protecting data privacy while simultaneously preserving data utility and ensuring a seamless user experience during interactions with CLLMs. The primary contributions of this work are summarized as follows:

- We discuss the challenges of balancing data privacy and data utility in interactions with CLLMs, which traditional encryption methods fail to overcome.
- We propose SE that protects data privacy while preserving data utility by transforming the user's input into semantically distinct yet logically equivalent contexts.
- Extensive experiments demonstrate that the proposed method effectively protects data privacy, preserves data utility, and maintains user experience.

Related Work

How to protect user privacy during interactions with CLLMs is gradually becoming a hot topic in current research. PrivacyRestore trains restoration vectors for each privacy span to alleviate insufficient privacy protection with performance degradation (Zeng et al. 2024). Some studies protect data privacy by modifying user input keywords using local differential privacy mechanisms; however, this often leads to a degradation in data utility (Li, Tan, and Liu 2023; Hoory et al. 2021; Du and Mi 2021). InferDPT (Tong et al. 2025) leverages differential privacy to safeguard data privacy while concurrently training a decoder to reconstruct the encrypted content. Although this approach achieves a degree of balance between data utility and privacy preservation, it inevitably incurs degradation in the logical structure of inputs.

Pretrained large models possess extensive prior knowledge and strong reasoning capabilities but face challenges in being deployed across a wide range of devices (Chen et al. 2024b,a). In contrast, although small models have limited performance, they are easier to deploy. Thus, some studies have attempted to develop various techniques to enable small models to acquire the capabilities of pretrained large models (Fang et al. 2025; Chen et al. 2025).

In this paper, we propose a plug-and-play framework where local small models transform the original input from users and reponses from CLLMs.

Methodology

As illustrated in Figure 2, the proposed Semantic Encryption (SE) framework enhances data privacy by transforming user inputs into a logically consistent but semantically different representation, thereby preserving both privacy and utility. The response from the CLLM is subsequently mapped back to the original semantic context with a Semantic Decoder, enabling seamless and effective user interaction. In order to support a smooth user experience, Semantic Distillation is introduced within the SE, allowing a lightweight local model to approximate the semantic transformation capabilities.

Semantic Encoding

The Semantic Encoder F_{SE} is implemented with a lightweight local model to transform the original input T_o into a logically consistent but semantically different representation $\widehat{T_o}$ in other contexts. However, effectively mapping original semantic contexts to alternative ones necessitates extensive prior knowledge and strong logical reasoning capabilities, which lightweight models often lack. To address this limitation, we propose Semantic Distillation that extracts the prior knowledge and semantic transformation capabilities of CLLMs and distills them into a lightweight model.

Specifically, to enable F_{SE} to acquire the semantic transformation capability of CLLMs F_{CLLM} across diverse contexts, we first employ a random number generator to produce random number list A with random lengths and values:

$$\begin{aligned} \mathbf{A} &= [a_1, a_2, \ldots, a_n], \\ &\text{where } n \sim \mathcal{U}(n_{\min}, n_{\max}), \\ &a_i \sim \mathcal{U}(v_{\min}, v_{\max}), \quad i = 1, 2, \ldots, n \end{aligned} \tag{1}$$

where the list length n follows a discrete uniform distribution $\mathcal{U}$ within the range $[n_{\min}, n_{\max}]$, each element x_i is randomly sampled from a given value range $[v_{\min}, v_{\max}]$, following a uniform distribution. The generated random number list, composed of various numerical combinations, is designed to stimulate the CLLM to produce a wide range of semantic contexts, thereby exploring more possibilities for semantic transformation.

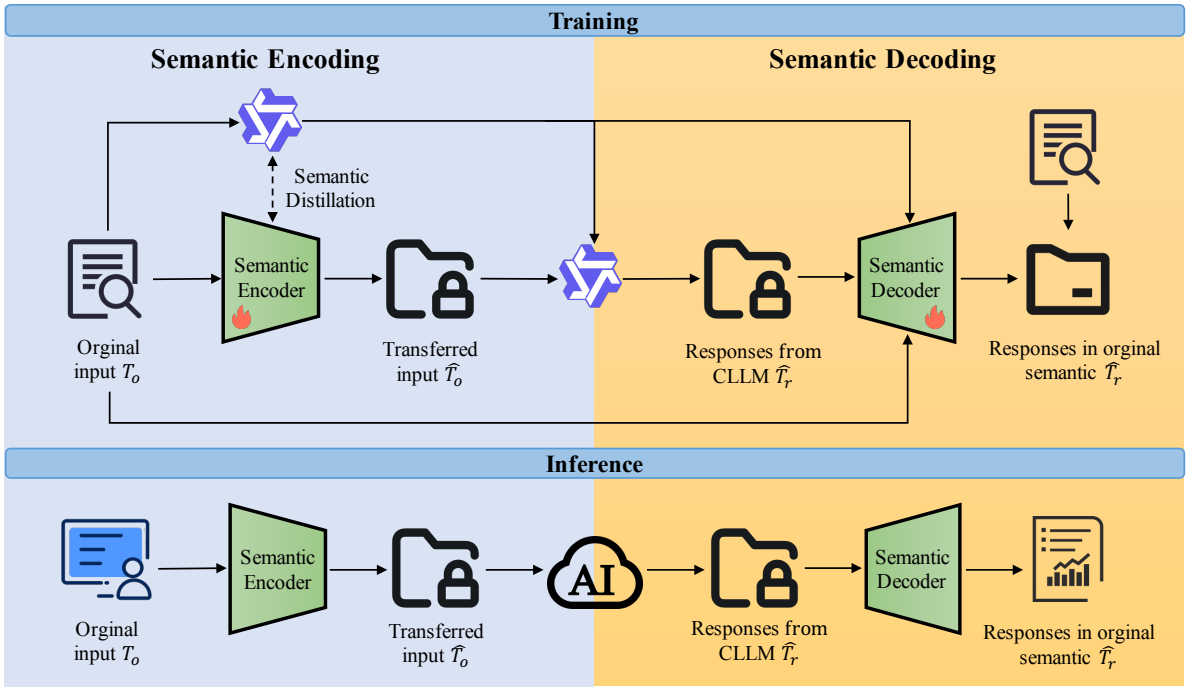

Figure 2: The proposed Semantic Encryption (SE) framework. SE consists of Semantic Encoding and Semantic Decoding, which focus on performance and user experience, respectively.

Then, F_{CLLM} can generate original input T_o based on the list A:

$$T_o = F_{CLLM}(A, P_0) \tag{2}$$

where P_0 is the prompt for generating T_o. Intuitively, the greater the diversity of T_o, the more beneficial it is for training a general-purpose semantic encoder. However, due to limited computational resources, this paper focuses solely on task-specific semantic encryption. That is, if existing datasets for different tasks are available, A in Equation 2 will no longer be necessary, as T_o in this case directly serves as the training data.

Based on the CLLM's extensive prior knowledge and powerful reasoning capabilities, the T_o can be transformed to alternative semantic contexts.

$$\widehat{T_o} = F_{CLLM}(T_o) \tag{3}$$

By constructing $(T_o, \widehat{T_o})$ pairs with Equation 3, we can extract the contexts knowledge and semantic transformation capability embedded in the CLLM. Subsequently, we fine-tune the lightweight local model with $(T_o, \widehat{T_o})$ pairs to enable effective transformation of user inputs by the following equation:

$$\min_{F_{SE}} \mathcal{L}(F_{SE}(T_o), \widehat{T_o}) \tag{4}$$

After obtaining the Semantic Encoder by Equation 4, we can send the transformed input to the CLLM to obtain corresponding response $\widehat{T_r}$:

$$\widehat{T_r} = F_{CLLM}(\widehat{T_o}) \tag{5}$$

The response $\widehat{T_r}$ from the CLLM remains within the transformed semantic context.

Semantic Decoding

Semantic Encoding can protect data privacy and preserves data utility by transforming the original input T_o to alternative semantic contexts $\widehat{T_o}$. However, this also causes the CLLM's responses $\widehat{T_r}$ to remain within the transformed semantic context, potentially impacting the user experience. To address this issue, we introduce Semantic Decoding into the SE framework. By training an efficient Semantic Decoder, SE enables the transformation of the CLLM's responses $\widehat{T_r}$ back to the original semantic context T_r.

Specifically, similar to Semantic Encoding, we provide the CLLM F_{CLLM} with the original input T_o and response $\widehat{T_r}$, enabling it to restore the response $\widehat{T_r}$ to the original semantic context. This process can be formally expressed as:

$$T_r = F_{CLLM}(T_o, \widehat{T_r}) \tag{6}$$

Subsequently, based on the $(T_o, \widehat{T_o}, \widehat{T_r}, T_r)$ quadruple, we train the Semantic Decoder F_{SD} by:

$$\min_{F_{SD}} \mathcal{L}(F_{SD}(T_o, \widehat{T_o}, \widehat{T_r}), T_r) \tag{7}$$

Similar to T_o in Equation 2, if datasets for different tasks are available, T_r in Equation 7 can be directly replaced by the labels provided in the datasets.

The whole training process of SE is summarized in Algorithm 1.

Can SE protect data privacy?

In this section, we demonstrate that semantic transformation achieved through SE can effectively protect data privacy in the sense of Shannon.

Algorithm 1: Training phase of Semantic Encryption.

Generate diverse context representations $\widehat{T_o}$ with CLLMs based on the training data T_o.
Train Semantic Encoder by Equation 4 with T_o and $\widehat{T_o}$.
Process and analyze $\widehat{T_o}$ with the CLLM to obtain the response $\widehat{T_r}$.
Treat labels T_r in the training data as response in the original semantic context.
Train Semantic Decoder by Equation 7 with T_o, $\widehat{T_o}$, $\widehat{T_r}$ and T_r.

Let M denote the orginal context, i.e. the set of all possible representations in context H. Any $m \in M$ represents one such description. N is the numerical data (e.g. measurements or statistics) that remain unchanged during transformation. C denotes the ciphertext space, i.e. the set of all possible representations in context B. We define a key space K, where each key $k \in K$ specifies a bijective "semantic mapping", i.e. $|K| = |C|$. Besides, we interpret Φ_k as the encryption function under key k, and denote:

$$E_k : M \to C \tag{8}$$

where k is chosen uniformly at random from $\mathcal{K}$.

For any fixed $m \in M$ and $c \in C$,

$$\begin{aligned} \Pr\big(M = m \mid C = c\big) &= \frac{\Pr(E_k(m) = c)\Pr(M = m)}{\Pr(C = c)} \\ &= \Pr(M = m) \end{aligned} \tag{9}$$

so the ciphertext C reveals no information about M. Furthermore, since N is independent of the mapping K and is revealed unchanged in this paper, the joint output (C, N) satisfies

$$I\big(M;\, C, N\big) = 0 \tag{10}$$

where $I(\cdot;\cdot)$ denotes mutual information. Thus, according to 10, the proposed SE can protect data privacy for the original content.

Experiments

In our experiments, we aim to (1) validate that SE can preserve data utility, (2) validate that SE can maintain the user experience, (3) validate that SE can protect data privacy, (4) demonstrate the specific workflow of SE through a case study,

Datasets and Settings

In this work, we conduct experiments on three mathematical reasoning datasets and one natural language inference (NLI) dataset:

Gsm8K (Cobbe et al. 2021): Gsm8K is a high-quality benchmark dataset comprising elementary school-level math word problems, with 7,473 training samples and 1,319 testing samples. Each problem necessitates multi-step arithmetic reasoning, typically involving between 2 and 8 steps.

Table 1: Data utility after user input encrypted by different privacy protection methods. The values in the table represent the accuracy of the CLLM in handling encrypted queries. For Gsm8K and Metamath, the values denote the probability that the CLLM generates the correct output. For ANLI, the values indicate the probability that the CLLM correctly identifies the underlying logical relationship.

Method	Gsm8K	Metamath	ANLI
SANTEXT	0.00%	0.08%	33.38%
SANTEXT+	1.29%	3.36%	33.47%
CUSTEXT	4.09%	17.78%	34.59%
CUSTEXT+	7.66%	23.64%	42.53%
HaS	1.97%	16.46%	45.59%
SE(Ours)	**83.02%**	**84.08%**	**55.78%**

OrcaMath (Mitra et al. 2024): OrcaMath is a synthetic dataset comprising mathematics problems reformulated using the GPT-4-Turbo model within the Agent-Instruct framework. From this dataset, 10,000 samples are selected for training and 5,000 samples for testing.

MetaMath (Yu et al. 2023): MetaMath is an augmented dataset derived from Gsm8K and MATH (Hendrycks et al. 2021) through techniques such as problem restatement and reverse reasoning, comprising a total of 395,000 samples. From this dataset, 10,000 samples are used for training and 5,000 for testing.

ANLI (Nie et al. 2020): ANLI is a NLI dataset constructed with multi-round human-and-model-in-the-loop adversarial training. It comprises three label categories: entailment, neutral, and contradiction. From this dataset, 15,000 samples are used for training, while 3,200 samples are used for testing.

We compare the proposed Semantic Encryption (SE) method with several differential privacy-based methods, including SANTEXT, SANTEXT+ (Yue et al. 2021), CUSTEXT, and CUSTEXT+ (Chen et al. 2023b), as well as HaS (Chen et al. 2023c) and InferDPT (Tong et al. 2025), which employs an encoder-decoder architecture. In all the following experiments, the CLLM is Qwen-Plus, and both the Semantic Encoder and Semantic Decoder in SE are Qwen3-0.6B (Yang et al. 2025). We train models with LoRA (Hu et al. 2022), the rank is 8, the learning rate is 2e-5, the batch size is 2. All experiments are conducted on a single NVIDIA RTX A6000 GPU. For more details on the experimental settings, please refer to the accompanying code. All prompts used in the experiments are provided in the Appendix A. The code and data for the proposed method are provided for research purpose [1].

Data Utility of Encrypted User Input for CLLMs

In this section, we evaluate the effectiveness of CLLM in handling queries encrypted by different methods across various tasks to verify that SE can preserve data utility.

As shown in Table 1, SE consistently achieves the best

[1]Code is included in the supplemental material. Code will be released upon the paper acceptance.

Table 2: Impact of various privacy protection methods on user experience. All methods encrypt the user queries prior to processing by the same CLLM. The responses returned by the CLLM are subsequently post-processed according to the specifications of each method. Finally, the differences between the user-received responses and manual annotations are analyzed to evaluate the impact of the various encryption methods on user experience. BLEU denotes the average score from BLEU-1 to BLEU-4. BERTScore (Zhang et al. 2019) quantifies the overall similarity between the responses and manual annotations.

Dataset	Method	BLEU↑	METEOR↑	R-1↑	R-2↑	R-L↑	BERTScore↑
Gsm8K	SANTEXT	0.0572	0.0509	0.0939	0.0012	0.0646	0.3676
	SANTEXT+	0.0728	0.0987	0.1323	0.0141	0.0899	0.4148
	CUSTEXT	0.0972	0.2061	0.2197	0.0486	0.1372	0.5166
	CUSTEXT+	0.1146	0.2793	0.2869	0.0799	0.1741	0.5758
	HaS	0.1299	0.2716	0.3713	0.1120	0.2385	0.6293
	InferDPT	0.1947	0.3687	0.4996	0.2139	0.3443	0.7149
	SE (ours)	**0.2294**	**0.4368**	**0.5380**	**0.2524**	**0.3682**	**0.7243**
OrcaMath	SANTEXT	0.0133	0.0379	0.1255	0.0023	0.0609	0.3827
	SANTEXT+	0.0452	0.0898	0.1859	0.0282	0.1019	0.4390
	CUSTEXT	0.1216	0.2266	0.3340	0.1032	0.1936	0.5757
	CUSTEXT+	0.1687	0.3091	0.4173	0.1590	0.2449	0.6354
	HaS	0.2019	0.3171	0.4933	0.2077	0.2850	0.6628
	InferDPT	0.3187	0.4443	0.6321	0.3628	0.4129	0.7516
	SE (ours)	**0.3638**	**0.4839**	**0.6771**	**0.4161**	**0.4614**	**0.7719**
Metamath	SANTEXT	0.0276	0.0460	0.1274	0.0026	0.0679	0.3732
	SANTEXT+	0.0712	0.1248	0.2147	0.0394	0.1184	0.4537
	CUSTEXT	0.1304	0.2347	0.3112	0.1005	0.1882	0.5622
	CUSTEXT+	0.1691	0.3184	0.3939	0.1523	0.2413	0.6271
	HaS	0.2288	0.3785	0.5176	0.2547	0.3335	0.6859
	InferDPT	0.3666	0.5056	0.6528	0.4202	0.4815	0.7754
	SE (ours)	**0.4272**	**0.5594**	**0.6942**	**0.4767**	**0.5268**	**0.7964**

performance across both mathematical reasoning and NLI datasets. In particular, on the two mathematical reasoning benchmarks, SE surpasses the second-best baseline, CUSTEXT+, by 75.36% on Gsm8K and 60.44% on MetaMath. This significant improvement can be attributed to SE's ability to transform the original input into semantically distinct yet logically equivalent contexts. In other words, SE effectively preserves critical components of mathematical reasoning, enabling the CLLM to accurately perform tasks based on encrypted inputs. In contrast, other methods employ random or rule-based substitutions of the user's original input, which undermines the preservation of the user's intention. On the other hand, although SE exhibits less pronounced advantages on the ANLI task compared to mathematical reasoning tasks, it still demonstrates its effectiveness on relatively simpler tasks. These results indicate that SE is more effective in contexts that rely on the reasoning capabilities.

User Experience with CLLM's Responses

In the Table 2, the encrypted inputs are initially processed by the CLLM. Subsequently, each method applies an additional post-processing step to the CLLM's output to generate the final responses presented to the user. By comparing the final responses with corresponding manual annotations, we can evaluate the impact of each method on user experience. It can be observed that methods incorporating a decoding process, such as HaS, InferDPT and the proposed SE, consistently outperform differential privacy-based methods, including SANTEXT, SANTEXT+, CUSTEXT, and CUSTEXT+. This is primarily because differential privacy introduces random noise to achieve encryption, which can cause a substantial semantic divergence between the encrypted and original inputs, thereby leading to a marked deviation between the CLLM's output and the intended result. In contrast, HaS, InferDPT and SE establish a mapping between the original and encrypted content, enabling the transformation of the CLLM's output back into a context that is more readily interpretable by users. InferDPT leverages a designed differential privacy strategy to encrypt the user's original input and subsequently trains a decoder to reconstruct it. However, InferDPT's encryption process inherently leads to the irreversible loss of certain useful information. Similarly, HaS relies exclusively on the substitution of sensitive terms, which often disregards the user's original intent and the logical structure of the input, potentially impairing the performance of the CLLM and adversely affecting the user experience. In comparison, SE transforms user input into a semantically similar context that preserves logical coherence, and leverages a semantic decoding module to reconstruct the CLLM's response. As a result, SE consistently outperforms all baselines across all datasets and evaluation metrics, which demonstrates that SE can effectively maintain user experience.

It is important to note that our implementation of Infer-

Table 3: Quantitative comparison of privacy protection results across different methods.

Dataset	Method	BLEU↓	METEOR↓	R-1↓	R-2↓
Gsm8K	SANTEXT	0.0492	0.0235	0.0228	0.0001
	SANTEXT+	0.2995	0.4919	0.5329	0.2958
	CUSTEXT	0.2153	0.3852	0.3782	0.2073
	CUSTEXT+	0.4347	0.6788	0.6273	0.4359
	HaS	0.6134	0.7941	0.8231	0.6664
	SE (Ours)	0.6109	0.7597	0.7954	0.6027
OrcaMath	SANTEXT	0.0529	0.0238	0.0251	0.0001
	SANTEXT+	0.3035	0.4841	0.5258	0.2990
	CUSTEXT	0.2170	0.3863	0.3761	0.2100
	CUSTEXT+	0.4383	0.6808	0.6216	0.4380
	HaS	0.6301	0.8043	0.8307	0.6830
	SE (Ours)	0.5951	0.7517	0.7918	0.5909
MetaMath	SANTEXT	0.0407	0.0207	0.0235	0.0014
	SANTEXT+	0.3287	0.5108	0.5614	0.3358
	CUSTEXT	0.3023	0.4514	0.4280	0.2744
	CUSTEXT+	0.5254	0.7325	0.6716	0.5102
	HaS	0.6955	0.8237	0.8481	0.7212
	SE (Ours)	0.7169	0.8167	0.8411	0.6929

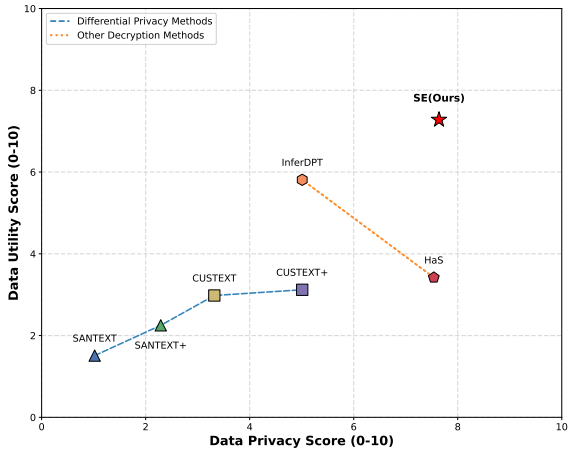

Figure 3: The average evaluation results of the LLM applied to various encryption methods across three mathematical reasoning datasets.

DPT adopts CUSTEXT+ for the encryption process. Consequently, its performance with respect to data utility and data privacy is consistent with that of CUSTEXT+. Therefore, InferDPT is only included in the Table 2.

Data Privacy of Encrypted User Input

In this section, we follow the prior work (Li et al. 2025; Chen et al. 2023c) to evaluate the privacy protection capabilities of different methods by assessing the similarity between the user's original input and its encrypted counterpart. The similarity is quantified using standard metrics such as BLEU and ROUGE. Intuitively, lower BLEU and ROUGE scores indicate stronger privacy protection. Related results are reported in Table 3.

The results presented in Table 3 indicate that traditional differential privacy-based methods, such as SANTEXT, SANTEXT+, CUSTEXT, and CUSTEXT+, are highly effective in preserving data privacy. This effectiveness is primarily attributed to the introduction of varying levels of noise into the original user inputs, whereby higher noise magnitudes lead to greater divergence between the encrypted content and the original data. However, as demonstrated by the results in Tables 1 and 2, the introduction of noise significantly compromises data utility, leading to notable degradation in both CLLM performance and user experience. In contrast, other decryption approaches such as HaS and the proposed SE, which adopt encoder–decoder architectures, offer a more favorable balance between privacy protection and data utility. This is achieved by selectively transforming only the sensitive components of user input, thereby avoiding unnecessary obfuscation of non-sensitive segments. Furthermore, compared to HaS—which operates by selectively masking identified sensitive terms—SE provides stronger privacy guarantees in most contexts. For instance, on the OrcaMath dataset under the R-2 metric, the similarity between SE-encrypted text and the original input is 0.0921 lower than that of HaS, indicating a greater degree of semantic transformation and, consequently, enhanced privacy protection.

Large Language Model Evaluation

In this section, we evaluate various encryption methods using the advanced LLM Qwen-Plus. Detailed evaluation metrics and prompts are provided in the Appendix A. The corresponding results are presented in Figure 3. Compared to the results presented in Table 3, this experiment places greater emphasis on privacy protection from the user's perspective,

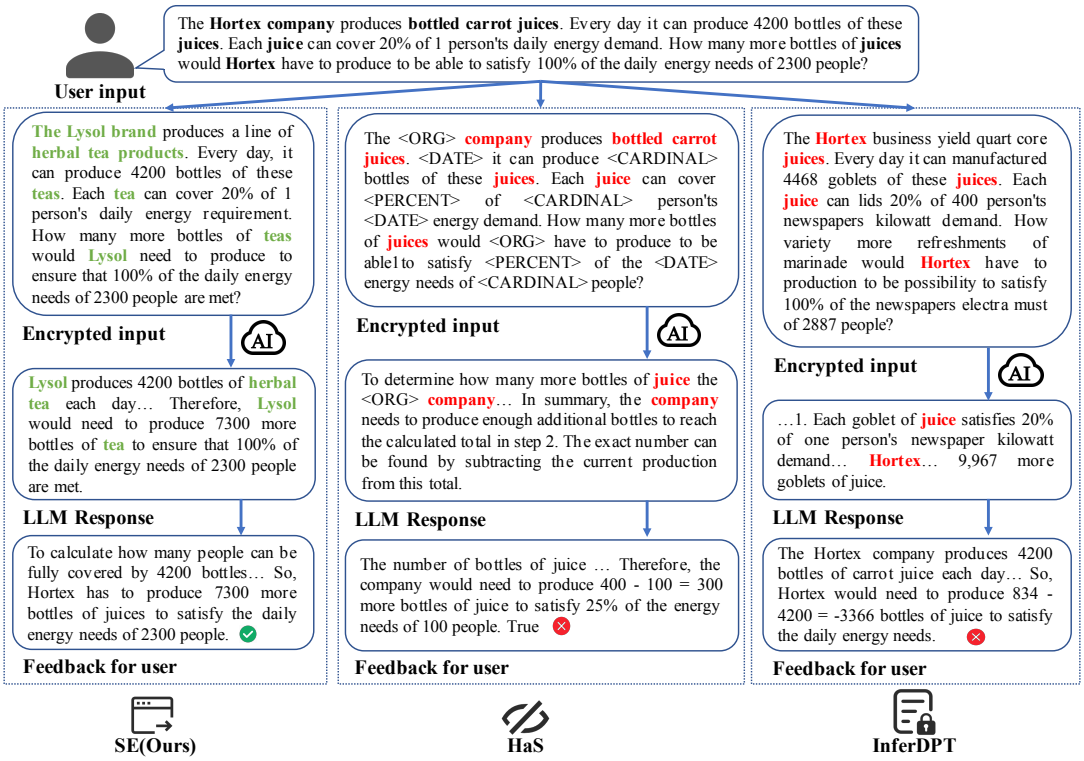

Figure 4: The case study for SE and Has. Bold black font denotes privacy-sensitive information requiring protection, red indicates instances of privacy leakage, and green signifies effective privacy protection.

including metrics such as Logical Structure Preservation. By effectively replacing irrelevant information while maintaining the original input's logical structure, the proposed SE achieves superior performance in data privacy protection. Furthermore, evaluations of response quality from the user's perspective such as Logical Reasonableness, further validate the suitability of the proposed method for interactive contexts involving CLLMs.

Case Study

We present the workflows of SE and the two best-performing baselines, HaS and InferDPT, in Figure 4. HaS protects private information through keyword substitution. However, its encrypted input reveals limitations in effectively securing all critical elements; for instance, terms like "juices" remain unencrypted, potentially disclosing that the input relates to a juice manufacturing company. As for InferDPT, it encrypts the user input with differential privacy. Nevertheless, its effectiveness in protecting data privacy is inconsistent—for instance, it fails to encrypt critical information such as company name in this case. Moreover, HaS and InferDPT compromises data utility, preventing the CLLM from responding based on the correct logical relationships. In contrast, SE treats numerical values as shareable information and transforms critical elements into alternative semantic contexts, thereby facilitating the effective execution of relevant analytical tasks. Besides, during the user feedback stage, SE restores the CLLM's response into the original semantic context and returns them to the user. Throughout the interaction with the CLLMs, users will remain unaware of the presence of the SE framework.

The Appendix B further reports the response accuracy of the three methods. Notably, SE substantially outperforms the baselines, particularly on the GSM8K dataset, achieving accuracy gains of 34.34% and 63.83% over InferDPT and HaS, respectively.

Conclusion

In light of the growing importance of interactions with CLLMs, traditional encryption techniques such as differential privacy offer protection for sensitive data but usually at the cost of data utility. The degradation of data utility impairs the ability of CLLMs to generate satisfactory responses to user queries, as traditional encryption techniques hinder the CLLMs' understanding of encrypted user inputs. To overcome these limitations, we introduce Semantic Encryption, a novel framework consists of Semantic Encoding and Semantic Decoding. For Semantic Encoding, a semantic encoder analyzes the user's original input and transforms it into an alternative semantic context that preserves the original logical structure. After the encrypted input is processed by the CLLM, a semantic decoder is employed to map the CLLM's response back to the original semantic context. This end-to-end process ensures data privacy throughout the interaction, while preserving high response quality and delivering a seamless user experience.

References

Chen, D.; Hu, Z.; Fan, P.; Zhuang, Y.; Li, Y.; Liu, Q.; Jiang, X.; and Xu, M. 2025. KKA: Improving Vision Anomaly Detection through Anomaly-related Knowledge from Large Language Models. *arXiv preprint arXiv:2502.14880*.

Chen, D.; Tang, S.; Shen, Z.; Wang, G.; Xiao, J.; Zhuang, Y.; and Yang, C. 2023a. FedAA: Using Non-sensitive Modalities to Improve Federated Learning while Preserving Image Privacy. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3796–3806.

Chen, D.; Zhang, S.; Zhuang, Y.; Tang, S.; Liu, Q.; Wang, H.; and Xu, M. 2024a. Improving large models with small models: Lower costs and better performance. *arXiv preprint arXiv:2406.15471*.

Chen, D.; Zhuang, Y.; Zhang, S.; Liu, J.; Dong, S.; and Tang, S. 2024b. Data shunt: Collaboration of small and large models for lower costs and better performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11249–11257.

Chen, S.; Mo, F.; Wang, Y.; Chen, C.; Nie, J.-Y.; Wang, C.; and Cui, J. 2023b. A Customized Text Sanitization Mechanism with Differential Privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, 5747–5758.

Chen, Y.; Li, T.; Liu, H.; and Yu, Y. 2023c. Hide and seek (has): A lightweight framework for prompt privacy protection. *arXiv preprint arXiv:2309.03057*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Dong, B.; Chen, D.; Wu, Y.; Tang, S.; and Zhuang, Y. 2024. FADngs: Federated learning for anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Du, J.; and Mi, H. 2021. Dp-fp: Differentially private forward propagation for large models. *arXiv preprint arXiv:2112.14430*.

Fang, L.; Yu, X.; Cai, J.; Chen, Y.; Wu, S.; Liu, Z.; Yang, Z.; Lu, H.; Gong, X.; Liu, Y.; et al. 2025. Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *arXiv preprint arXiv:2504.14772*.

Feretzakis, G.; Papaspyridis, K.; Gkoulalas-Divanis, A.; and Verykios, V. S. 2024. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. *Information*, 15(11): 697.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Hoory, S.; Feder, A.; Tendler, A.; Erell, S.; Peled-Cohen, A.; Laish, I.; Nakhost, H.; Stemmer, U.; Benjamini, A.; Hassidim, A.; et al. 2021. Learning and evaluating a differentially private pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1178–1189.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Li, G.; Zhang, Y.; Wang, Y.; Yan, S.; Wang, L.; and Wei, T. 2025. PRIV-QA: Privacy-Preserving Question Answering for Cloud Large Language Models. *arXiv preprint arXiv:2502.13564*.

Li, Y.; Tan, Z.; and Liu, Y. 2023. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212*.

Mitra, A.; Khanpour, H.; Rosset, C.; and Awadallah, A. 2024. Orca-Math: Unlocking the potential of SLMs in Grade School Math. arXiv:2402.14830.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901.

Tong, M.; Chen, K.; Zhang, J.; Qi, Y.; Zhang, W.; Yu, N.; Zhang, T.; and Zhang, Z. 2025. Inferdpt: Privacy-preserving inference for black-box large language models. *IEEE Transactions on Dependable and Secure Computing*.

Wang, J. G.; Wang, J.; Li, M.; and Neel, S. 2024. Pandora's White-Box: Increased Training Data Leakage in Open LLMs. *arXiv e-prints*, arXiv–2402.

Wu, Y.; Li, Z.; Zhang, J. M.; and Liu, Y. 2024. Condefects: A complementary dataset to address the data leakage concern for llm-based fault localization and program repair. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 642–646.

Yan, B.; Li, K.; Xu, M.; Dong, Y.; Zhang, Y.; Ren, Z.; and Cheng, X. 2024. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Yang, S.; Chiang, W.-L.; Zheng, L.; Gonzalez, J. E.; and Stoica, I. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*.

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*, 100211.

Yu, L.; Jiang, W.; Shi, H.; Yu, J.; Liu, Z.; Zhang, Y.; Kwok, J. T.; Li, Z.; Weller, A.; and Liu, W. 2023. MetaMath: Bootstrap Your Own Mathematical Questions for Large Language Models. *arXiv preprint arXiv:2309.12284*.

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

Zeng, Z.; Wang, J.; Yang, J.; Lu, Z.; Zhuang, H.; and Chen, C. 2024. PrivacyRestore: Privacy-Preserving Inference in Large Language Models via Privacy Removal and Restoration. *arXiv preprint arXiv:2406.01394*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.