JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 11, pp. 91-98, November 2025
https://doi.org/10.9708/jksci.2025.30.11.091

SafeDP-Rewrite: Differentially Private Text Rewriting with Black-Box Access to Large Language Models

Jong Wook Kim*

*Professor, Dept. of Computer Science, Sangmyung University, Seoul, Korea

[Abstract]

Text data is a critical resource in modern machine learning applications but often contains sensitive information, creating risks of privacy leakage when shared. Differential privacy (DP) provides a theoretical guarantee to prevent such leakage during data sharing, and recent work has explored its application to text rewriting using large language models (LLMs). However, most existing approaches assume a white-box setting with access to internal LLM model structures, making them impractical in real-world scenarios where only black-box API access is available. To address this limitation, we propose SafeDP-Rewrite, a DP-based text rewriting method that operates entirely in the black-box setting of LLMs. The proposed method generates diverse candidate sentences through random masking and applies the exponential mechanism to ensure DP in the final output. SafeDP-Rewrite requires neither additional training nor access to internal model information, making it simple and practical to deploy. Experiments on real-world datasets demonstrate that the proposed method preserves semantic fidelity and fluency while simultaneously achieving both privacy protection and utility.

▸ **Key words**: Differential privacy, Text rewriting, Black-box large language models, Data privacy

[요 약]

텍스트 데이터는 현대 기계학습 응용에서 핵심적 자원이지만, 민감한 정보를 포함하는 경우가 많아 공유 과정에서 개인정보 유출 위험이 발생할 수 있다. 차분 프라이버시(differential privacy, DP)는 데이터 공유 과정에서 프라이버시 유출 방지를 이론적으로 보장하며, 최근에는 대규모 언어모델(LLM) 기반 텍스트 재작성 기법에도 활용되고 있다. 그러나 기존 연구들은 LLM 모델 내부 구조에 접근할 수 있는 화이트박스 환경을 전제로 하여, 실제와 같은 블랙박스 API 환경에서는 활용이 어렵다는 한계가 있다. 본 논문에서는 이러한 제약을 해결하기 위해 SafeDP-Rewrite라는 LLM 블랙박스 기반 DP 텍스트 재작성 기법을 제안한다. 제안 방식은 무작위 마스킹을 통해 다양한 후보 텍스트를 생성하고, 지수 메커니즘을 적용하여 DP를 보장하는 최종 텍스트를 선택한다. SafeDP-Rewrite는 추가 학습이나 내부 정보 접근이 필요 없으며, 단순하고 실용적인 활용이 가능하다. 실데이터셋을 이용한 실험 결과, 제안 기법은 문장의 의미와 유창성을 유지하면서도 프라이버시와 유용성을 동시에 확보할 수 있음을 보였다.

▸ **주제어**: 차분 프라이버시, 텍스트 재작성, 블랙박스 대규모 언어 모델, 데이터 프라이버시

• First Author: Jong Wook Kim, Corresponding Author: Jong Wook Kim
*Jong Wook Kim (jkim@smu.ac.kr), Dept. of Computer Science, Sangmyung University
• Received: 2025. 09. 22, Revised: 2025. 10. 20, Accepted: 2025. 11. 07.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

인공지능 기술의 비약적인 발전으로 대규모 데이터 활용 수요가 꾸준히 증가하고 있다. 특히 자연어처리 분야에서는 방대한 양의 텍스트가 모델 학습과 평가에 핵심 자원으로 활용되고 있다. 그러나 의료, 법률, 금융과 같이 민감한 영역에서 수집되는 텍스트에는 개인의 민감한 정보가 포함되는 경우가 많아, 이를 연구나 응용 개발 목적으로 공유할 경우 심각한 개인정보 유출 위험이 발생할 수 있다 [1,2,3]. 이에 따라 개인정보를 보호하면서도 텍스트 데이터를 안전하게 활용할 수 있는 방안이 요구되고 있다.

이러한 문제를 해결하기 위한 방안으로 차분 프라이버시(differential privacy, DP)가 주목받고 있다. DP는 민감한 데이터가 외부에 노출되지 않도록 보장하는 개인정보 보호 기법이다 [4]. 초기 연구들은 텍스트 데이터에 DP를 적용하기 위해 토큰 수준에서 단어 임베딩에 노이즈를 주입하거나, DP로 보정된 확률에 따라 임의의 토큰으로 대체하는 방식을 사용하였다. 그러나 이러한 접근은 비문법적이거나 의미가 왜곡된 문장을 생성할 가능성이 있다 [5,6,7]. 최근에는 원문의 의미와 문장 유창성을 유지하면서도 DP 제약을 만족하도록 문장을 재작성하는 방식이 제안되고 있다 [8].

최근 대규모 언어모델(large language model, LLM) [9,10]의 확산과 함께, LLM의 생성 능력을 활용하여 민감한 텍스트를 DP 방식으로 비식별화하려는 연구가 이루어지고 있다. 그러나 기존 연구의 대부분은 모델의 내부 연산 결과나 파라미터에 직접 접근할 수 있는 화이트박스(white-box) 환경을 전제로 한다 [11,12,13]. 이러한 방식은 생성 과정을 세부적으로 조정할 수 있다는 장점이 있으나, 모델을 로컬 환경에서 직접 구동해야 하므로 자원 제약이 있는 환경에서는 적용이 어렵다는 한계가 있다.

반면, 실제 환경에서는 대부분의 사용자가 상용 API를 통해 LLM을 이용하며, 이 경우 모델 내부 정보에 접근할 수 없고 입력과 출력만 주고받는 블랙박스(black-box) 환경에 해당한다 [14,15,16]. 이러한 제약을 고려할 때, 모델 내부 접근 없이 API 기반의 입출력만으로 동작하는 DP 텍스트 재작성 기법의 필요성이 점점 커지고 있다. 이러한 방식은 로컬에서 모델을 직접 구동하거나 추가 학습을 수행하기 어려운 환경에서도 DP 기반 텍스트 비식별화를 가능하게 하여 활용 범위를 크게 확장할 수 있다.

본 논문에서는 이러한 요구를 충족하기 위해 SafeDP-Rewrite라는 블랙박스 기반 DP 텍스트 재작성 기법을 제안한다. 제안 기법은 모델 내부 접근이나 추가 학습, 로컬 추론 없이도 동작하므로 경량적이며 실용적이다. 구체적으로, 입력 문장을 무작위 마스킹 방식으로 변형한 뒤 블랙박스 LLM을 통해 여러 후보 문장을 생성하고, 이후 지수 메커니즘(exponential mechanism)을 적용하여 최종 출력을 선택한다. 이러한 설계를 통해 SafeDP-Rewrite는 ε-DP 보장을 제공하면서도 의미적 충실성과 문장 유창성을 유지할 수 있다.

본 논문의 주요 기여는 다음과 같다.

- 본 연구에서는 모델 내부 정보에 접근하지 않고 블랙박스 방식으로 LLM을 활용하는 새로운 DP 기반 텍스트 변조 기법인 SafeDP-Rewrite를 제안한다.
- SafeDP-Rewrite는 블랙박스 LLM API와의 입출력만을 활용한다. 입력 문장을 무작위 마스킹 방식으로 변형해 후보 문장을 생성한 뒤, 지수 메커니즘을 적용하여 모델 내부 연산 정보 없이도 DP 보장을 만족하는 최종 출력을 선택한다.
- 제안 기법의 성능을 검증하기 위해 다양한 실험을 수행하였다. 그 결과 SafeDP-Rewrite는 의미 보존과 문장 유창성 측면에서 우수한 품질을 유지하면서도 DP를 만족하였으며, API 기반 접근만 가능한 환경에서도 안정적으로 동작함을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하고, 3장에서는 본 연구에서 다루는 문제를 정의한다. 4장에서는 제안 기법을 상세히 설명하며, 5장에서는 실제 데이터셋을 이용한 실험을 통해 제안 기법의 성능을 평가한다. 마지막으로 6장에서 결론을 제시한다.

II. Related Work

기존 연구에서는 텍스트 데이터에 DP를 적용하기 위해 주로 단어 수준의 변형 기법이 제안되었다. 초기 접근법은 사전 학습된 임베딩 모델을 활용하여 각 단어를 벡터 공간에 매핑한 뒤 최근접 이웃을 기반으로 대체하거나, Mahalanobis 거리를 이용해 단어 임베딩에 타원형 노이즈를 주입하는 방식(즉, 확률분포에서 샘플한 임의의 값을 값을 임베딩/확률에 더해 민감도를 낮추는 방식)을 사용하였다 [5,6]. 이후에는 지수 메커니즘을 변형하여 인접 후보 가운데 단어를 선택하는 절단 지수 메커니즘(truncated exponential mechanism) [7], 무작위 클러스터링을 통해 임베딩을 변형하는 방법 [17], 그리고 지역 차분 프라이버시(local differential privacy, LDP) 메커니즘을 적용해 텍스트로 변환하는 방법 등이 제안되었다 [18]. 최근에는

토큰별 후보 집합을 구성한 뒤 지수 메커니즘을 적용하여 기존의 거리 기반 가정을 피하면서 각 토큰을 개별적으로 처리하는 방법도 연구되고 있다 [19].

문장 수준 접근법은 단어 단위의 변형이 아니라 문장 전체를 재작성하여 의미적 일관성과 문맥적 유창성을 보존하는 데 초점을 둔다. 예를 들어, 문서 임베딩을 기반으로 문장 수준에서 로컬 DP를 보장하는 SentDP가 제안되었으며, 이는 문서 내 특정 문장을 다른 문장으로 대체하더라도 임베딩 공간에서 큰 변화가 발생하지 않도록 설계되었다 [8]. 또한 문장 임베딩에 Mahalanobis 노이즈를 적응적으로 적용하여 프라이버시와 유틸리티 간의 절충을 개선하는 방법도 제안되었다 [20].

최근에는 LLM을 활용하여 문장 수준에서 DP 기반 텍스트 재작성을 수행하려는 연구가 활발히 이루어지고 있다. 한 연구에서는 파인튜닝된 트랜스포머 모델을 이용해 문장을 재작성함으로써 단어 수준 접근법의 의미적·구문적 한계를 보완하면서 DP를 보장하는 방법을 제안하였다 [11]. 또 다른 연구에서는 LLM의 제로샷 프롬프트를 활용해 LDP를 적용하는 DP-Prompt 기법을 제안하였다 [12]. 최근에는 마스킹 언어모델에 화이트박스 방식으로 접근하여 지수 메커니즘을 이용해 토큰 단위 비식별화를 수행하는 방법도 제시되었다 (여기서 마스킹은 특정 토큰을 MASK 등으로 가려 모델이 문맥으로 대체하도록 유도하는 전처리를 의미한다.) [13]. 그러나 이들 방법은 모두 모델 내부 정보에 접근할 수 있어야 한다는 제약이 있다.

이에 비해 본 연구에서 제안하는 SafeDP-Rewrite는 입력과 출력만을 활용하는 블랙박스 방식으로 동작하며, 모델 내부 정보에 전혀 의존하지 않는다. 이러한 차별성을 통해 기존 화이트박스 기반 접근의 한계를 극복하고, 실제 환경에서 활용 가능한 실용적인 DP 기반 텍스트 재작성 기법을 제공한다.

III. Background and Problem Definition

3.1. Background

DP는 확률적 알고리즘의 출력이 단일 데이터 포함 여부에 따라 크게 달라지지 않도록 보장하는 프라이버시 보존 프레임워크이다. 확률적 메커니즘 A가 ε-DP를 만족한다는 것은, 임의의 인접 입력 x, x'와 임의의 부분집합 $S \subseteq X$에 대하여 다음이 성립함을 의미한다.

$$\Pr[A(x) \in S] \leq e^{\epsilon} \cdot \Pr[A(x') \in S].$$

여기서 인접 입력 x, x'은 하나의 데이터 항목만 다른 경우를 가리킨다. 매개변수 ε은 프라이버시와 유용성 간의 균형을 결정한다. 작은 값은 강력한 보호를 보장하지만 결과의 활용도를 낮추고, 큰 값은 보호 수준을 완화하는 대신 더 높은 유용성을 제공한다.

DP를 만족하는 다양한 기법이 존재하며, 그중 지수 메커니즘은 이산적인 후보 집합에서 유틸리티 함수를 기반으로 출력을 선택하면서도 ε-DP를 보장하는 방법이다. 출력 공간이 수치형이 아니어서 단순 노이즈 주입 기법이 적용되기 어려운 경우 특히 유용하다. 후보 집합 C와 유틸리티 함수 $u(x,c)$가 주어졌을 때, 지수 메커니즘은 각 후보를 다음 확률에 비례하여 선택한다.

$$\Pr[A(x) = c] \propto \exp\left(\frac{\epsilon \cdot u(x,c)}{2\triangle u}\right),$$

여기서 $\triangle u$는 인접 입력에 따른 유틸리티 함수의 최대 변화량을 의미한다.

3.2. Problem Defnition

민감한 입력 텍스트를 T개의 토큰으로 이루어진 $x = (x_1, ..., x_T)$라 하자. 본 연구에서는 x를 재작성하여 $\tilde{x}$를 생성하는 무작위 메커니즘 M을 설계하는 것이다. 이때 $\tilde{x}$는 다음 세 가지 조건을 만족해야 한다. 첫째, 원문 x의 의미를 충실히 보존하면서 문법적으로 자연스러워야 한다. 둘째, 생성 과정은 원문 x에 대하여 ε-DP를 보장해야 한다. 셋째, 메커니즘은 LLM의 내부 정보(logit, 가중치 등)에 의존하지 않고 블랙박스 API 호출만으로 동작해야 한다. 이를 수식으로 표현하면 다음과 같다.

$$M(x) = \tilde{x}, \ M\text{은 } \epsilon - DP\text{를 만족한다.}$$

따라서 본 논문의 목표는 원문의 의미를 보존하면서 ε-DP를 만족하고, 블랙박스 LLM 환경에서도 실용적으로 동작할 수 있는 프라이버시 보존 텍스트 재작성 메커니즘을 설계하는 데 있다. 이와 같이 생성된 텍스트는 프라이버시 유출의 위험 없이 외부에 안전하게 공유될 수 있다.

본 연구는 두 가지 공격자를 고려한다. 첫째, 출력 기반 공격자(output-based attacker)는 최종 결과만을 관찰할 수 있으며 임의의 보조 정보를 가질 수 있다. 본 논문의 프라이버시 보장은 이 공격자를 기준으로 정의된다. 둘째, LLM 제공자(LLM provider)는 API 요청에 포함된 모든 프롬프트를 관찰할 수 있다고 가정한다. 다만, 제공자는 Semi-Honest한 공격자로 설정되며, API 처리는 올바르게 수행한다. 이러한 가정은 실제 배포 환경을 반영한다. 주요 LLM 제공자(OpenAI, Google 등)는 엔터프라이즈 환경에서 입력을 저장하거나 학습에 활용하지 않겠다는 계약적 보장을 제공한다. 따라서 본 논문은 제공자를 Semi-Honest한 존재로 간주하고, 주된 보호 대상은 출력 기반 공격자에 대한 프라이버시 보장으로 설정한다.

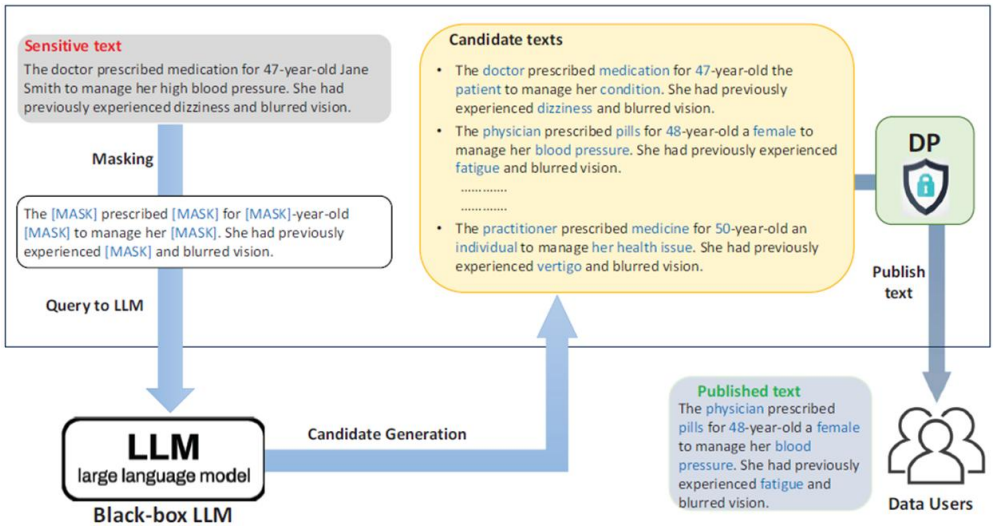


Fig. 1. Overview of the SafeDP-Rewrite process: (a) masking and candidate generation and (b) DP-based selection.

IV. Proposed Method

본 장에서는 제안 기법인 SafeDP-Rewrite를 설명한다. 그림 1은 전체 절차를 나타내며, 과정은 두 단계로 구성된다. 첫 번째 단계에서는 입력 문장의 일부 토큰을 베르누이 샘플링을 기반으로 마스킹한 뒤 이를 LLM에 입력하여 다수의 후보 문장을 생성한다. 두 번째 단계에서는 생성된 후보 집합에 지수 메커니즘을 적용하여 최종 출력을 선택한다. 이때 후보별 유틸리티 함수에 따라 확률적으로 선택이 이루어지며, 이 과정에서 ε-DP가 보장된다.

4.1. Candidate Generation

SafeDP-Rewrite의 첫 번째 단계는 입력 텍스트 $x=(x_1,...,x_T)$에 토큰 단위 무작위 마스킹을 적용하는 것이다. 이 과정은 두 가지 목적을 가진다. 첫째, 입력 전체 대신 부분적으로 마스킹된 문장을 LLM에 전달함으로써 Semi-Honest 제공자에게 노출되는 민감 정보를 최소화한다. 이는 DP 보장에는 포함되지 않지만 실제 환경에서 유용한 추가 보호 수단이 된다. 둘째, 마스킹은 후보 문장의 다양성을 확보하여 선택 단계에서 의미 있는 무작위성을 가능하게 한다. 모든 후보가 원문과 지나치게 유사하다면 선택 과정은 사실상 결정론적으로 작동하여 보호 효과가 약화된다. 반대로 후보들이 서로 다른 의미적 거리를 가진다면, 지수 메커니즘이 이를 활용해 더욱 강력한 프라이버시 보장을 제공할 수 있다.

이를 위해 먼저 항상 마스킹해야 하는 고위험 토큰 집합 $E\subseteq\{1,...,T\}$을 정의한다. 그 외의 위치 $i\notin E$에 대해서는 독립적인 베르누이 확률변수 $R_i\sim Bernoulli(p)$를 사용해 마스킹 여부를 결정한다. 마스킹 비율은 $p\in(0,1)$로 설정되며, 최종 마스킹 입력 $x^{mask}=(z_1,...,z_T)$는 다음과 같이 정의된다.

$$z_i=\begin{cases}[MASK] & \text{if } i\in E \text{ or } R_i=1,\\ x_i & otherwise.\end{cases}$$

마스킹된 입력은 블랙박스 LLM에 전달되어 후보 집합 $Y=\{y^{(1)},...,y^{(k)}\}$을 생성한다. 이후 각 후보와 원문 간의 의미적 유사도를 측정하고, 이를 기반으로 후보 집합의 의미적 분산도를 정량화한다. 이 분산도는 다음과 같이 정의된다.

$$D(Y)=\frac{1}{k}\sum_{j=1}^{k}\left(u_j-\overline{u}\right)$$

여기서 u_j는 j번째 후보 텍스트와 입력 텍스트 사이의 의미적 유사도를 나타내며, $\overline{u}$는 모든 후보 유사도의 평균값을 의미한다.

그림 2는 SafeDP-Rewrite의 후보 집합 생성 알고리즘을 나타낸다. 이 알고리즘은 후보 집합의 의미적 분산도가 사전에 정의된 임계값 δ 이상일 때만 해당 집합을 채택한다. 이를 위해 적응형 마스킹 기법을 적용한다. 먼저 초기 확률 p로 입력을 마스킹하고, 이를 LLM에 전달하여 후보를 생성한 뒤 의미적 분산도를 계산한다. 계산된 값이 임계치에 도달하지 못하면, 남아 있는 토큰에 대해 $Bernoulli(p)$ 샘플링을 추가로 적용하여 마스킹 비율을 높이고 과정을 반복한다. 이 절차는 분산도 조건을 충족하거나 최대 반복 횟수에 도달할 때까지 반복된다.

Require: Input text $x = (x_1, \ldots, x_T)$, masking probability p, diversity threshold δ, max iterations K
Initialize mask vector $m \in \{0,1\}^T$ using Bernoulli(p)
Initialize Y as an empty set
for $t = 1$ to K **do**
  Generate x^{mask} by applying m to x
  Query LLM with x^{mask} to obtain candidates Y_t
  $Y \leftarrow Y \cup Y_t$
  Compute diversity $D(Y)$
  if $D(Y) \geq \delta$ **then**
    return Y
  Apply additional Bernoulli(p) masking to unmasked tokens in m
return Y

Fig. 2. Candidate generation algorithm

4.2. Differentially Private Selection Mechanism

SafeDP-Rewrite의 두 번째 단계에서는 ε-DP를 만족하는 최종 출력을 선택한다. 이를 위해 지수 메커니즘을 활용하며, 각 후보가 원문과 얼마나 의미적으로 가까운지를 나타내는 유틸리티 지표(utility score)에 따라 확률적으로 출력을 결정한다. 후보 y의 유틸리티는 지표는 $u(y,x) \in [0,1]$로 정의되며, 이는 임베딩 기반의 코사인 유사도에 해당한다. 유사도가 [0,1] 구간에 제한되므로, 전역 민감도를 $\triangle u = 1$로 설정하면 지수 메커니즘이 최악의 경우에도 ε-DP를 보장한다.

지수 메커니즘은 DP에서 널리 활용되지만, LLM이 생성한 후보 집합에 직접 적용하면 한계가 드러난다. 하나의 질의에서 여러 출력을 요청할 경우, LLM은 의미는 거의 동일하지만 표현만 약간 다른 문장을 자주 생성한다. 예를 들어 어순 변화, 구두점 차이, 단어 치환 등은 사실상 같은 문장으로 볼 수 있다. 이러한 중복 후보가 많아지면 지수 메커니즘의 확률 분포가 왜곡되어, 의미적으로 차별성이 큰 후보가 선택될 가능성이 낮아진다.

이 문제를 해결하기 위해 본 연구에서는 유사도 절단 지수 메커니즘(similarity-truncated exponential bechanism, STEM)을 제안한다. STEM은 샘플링 전에 후보 집합을 정제하여 중복을 제거하는 절차이며, DP 보장에는 영향을 주지 않는다. 구체적으로, 먼저 후보 간 의미 유사도를 계산하고, 유사도가 임계값 τ 이상인 후보들을 하나의 그룹으로 묶는다. 이후 각 그룹에서 유틸리티 지표가 가장 높은 후보만 남겨 정제된 후보 집합 $\hat{Y}$를 구한다.

최종적으로 지수 메커니즘은 정제된 후보 집합 $\hat{Y}$에 적용된다. 각 후보 $y^{(j)} \in \hat{Y}$는 다음 확률로 선택된다.

$$\Pr[J = j \mid x] = \frac{\exp\left(\frac{\epsilon}{2} u(y^{(j)}, x)\right)}{\sum_{y' \in Y} \exp\left(\frac{\epsilon}{2} u(y', x)\right)}.$$

여기서 ε은 프라이버시 예산을 의미하며, 선택된 후보 $y^{(j)}$가 최종 재작성 결과로 반환된다. 이 과정을 통해 SafeDP-Rewrite는 후보 중복으로 인한 왜곡을 줄이면서도 ε-DP를 보장한다.

V. Experiments and Results

5.1. Experiment Setup

제안 기법의 성능을 평가하기 위해 두 가지 실데이터셋을 활용하였다. 먼저, MedQuAD 데이터셋[21]은 건강 관련 질문과 답변으로 구성되어 있으며, 이 중 500개의 질문-답변 쌍을 추출하여 사용하였다. 다음으로, IMDB 영화 리뷰 데이터셋[22]은 긍정과 부정으로 분류된 리뷰로 이루어진 감성 분석용 자료이며, 이 중 500개의 리뷰를 추출하여 실험에 사용하였다.

실험에서는 다음 세 가지 방법과 성능을 비교하였다.

- WordPerturb (WP) [5]: 단어 수준에서 프라이버시 보호를 수행하는 기존 기법에 해당한다.
- SafeDP-Rewrite without STEM (SDP-NS): 제안 기법에서 STEM 과정을 제외하고, 생성된 모든 후보에 지수 메커니즘을 직접 적용한 방법이다.
- SafeDP-Rewrite with STEM (SDP-S): 본 논문에서 제시하는 최종 방법에 해당한다.

성능 평가는 원문 텍스트와 최종 출력 간의 의미적 유사도로 측정하였다. 유사도는 사전 학습된 Sentence-BERT [23] 임베딩을 이용해 계산하였다. 알고리즘 구현은 API 기반 Gemini-2.0-Flash 모델을 사용하였다. 마스킹 확률 p는 {0.1, 0.2, 0.3, 0.4}에서 변화시켰으며, 프라이버시 예산 ε은 {0.5, 1.0, 2.0, 3.0}으로 설정하였다. 그림 2의 알고리즘에서는 최대 반복 횟수 $K = 5$, 다양성 임계값 $\delta = 0.35$를 사용하였다.

5.2. Experimental Results

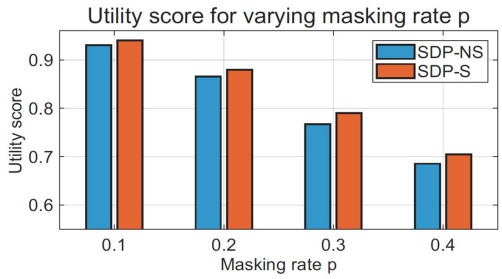
먼저, 마스킹 확률 p를 변화시키며 생성된 후보의 유틸리티 지표를 평가하였다. 이는 SafeDP-Rewrite의 후보 생성 과정이 입력 변형 수준에 따라 어떻게 달라지는지를 확인하기 위함이다. 표 1의 결과에서 보듯, 두 데이터셋 모두에서 p가 커질수록 유틸리티는 점진적으로 감소하였

다. 이는 마스킹 비율 증가로 문맥 정보가 줄어들고, 그 결과 LLM이 의미적으로 더 다양한 문장을 생성하게 되기 때문이다. 이러한 다양성은 DP 선택 단계에서 무작위성을 확보하는 데 기여한다.

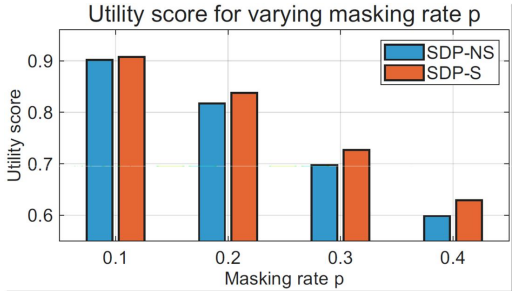
Table 1. Utility score of candidates for varying masking rate p.

Dataset	p=0.1	p=0.2	p=0.3	p=0.4
MedQuAD	0.9079	0.8296	0.7187	0.6856
IMDB	0.9359	0.8743	0.6949	0.6068

세부적으로 보면, MedQuAD 데이터셋에서는 평균 유틸리티가 p=0.1에서 0.9079였으나 p=0.4에서 0.6856으로 약 24.5% 감소하였다. IMDB 데이터셋에서는 감소폭이 더 커서 0.9359에서 0.6068로 줄어 약 35.2% 감소하였다. 이 결과는 p가 후보 다양성을 조절하는 핵심 매개변수임을 보여준다. 낮은 마스킹 비율(p=0.1-0.2)에서는 원문 의미가 잘 보존되어 유용성이 중요한 상황에 적합하다. 반면 높은 마스킹 비율($p \geq 0.3$)에서는 유틸리티가 다소 감소하지만, 후보의 의미적 범위가 넓어져 DP 기반 무작위 선택의 효과를 강화할 수 있다.



(a) MedQuAD

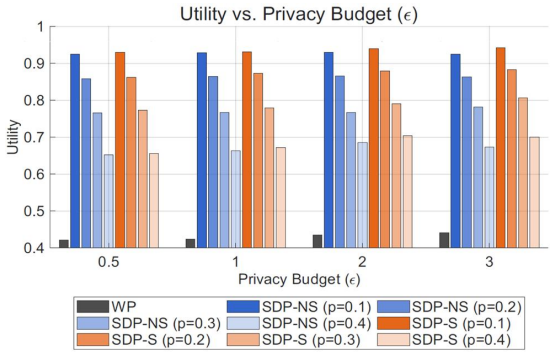


(b) IMDB

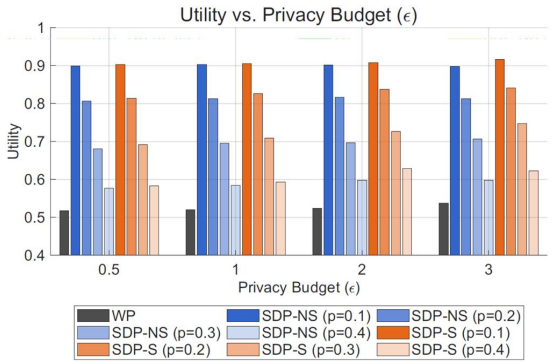
Fig. 3. Utility scores of SDP-NS and SDP-S for different masking rates p under a fixed privacy budget ε=2.0.

STEM 기법의 효과를 검증하기 위해, 그림 3에서는 프라이버시 예산 ε=2.0으로 고정한 상태에서 마스킹 비율 $p \in \{0.1, 0.2, 0.3, 0.4\}$에 따른 SDP-S와 SDP-NS의 유틸리티를 비교하였다. SDP-NS는 생성된 모든 후보에 지수 메커니즘을 직접 적용하는 방식이며, SDP-S는 STEM으로 후보를 정제한 뒤 동일한 메커니즘을 적용한다. 두 데이터셋 모두에서 마스킹 비율 p가 커질수록 유틸리티는 점차 감소하였으며, 이는 변형 강도가 높아질수록 원문의 의미적 충실성이 약화된다는 예상과 일치한다. 특히 모든 p 구간에서 SDP-S가 SDP-NS보다 일관되게 더 높은 유틸리티를 보여, STEM의 효과가 안정적으로 나타남을 확인할 수 있다.

그림 3의 실험 결과 차이는 STEM이 후보 집합에서 불필요하게 겹치는 문장을 걸러내는 데서 비롯된다. 중복을 제거하지 않으면 비슷한 품질 낮은 후보들이 확률을 나눠 가지면서, 오히려 의미가 충실한 문장이 선택될 기회를 줄이게 된다. STEM은 각 그룹에서 대표성 있는 후보만 남기기 때문에 다양성과 품질을 동시에 확보할 수 있으며, 그 결과 지수 메커니즘이 더 유용한 문장에 집중할 수 있도록 해준다. 특히 마스킹 비율이 커질수록 잡음이 많은 후보가 생성되므로, STEM의 효과는 더욱 두드러진다.



(a) MedQuAD



(b) IMDB

Fig. 4. Utility scores for WP, SDP-NS, and SDP-S across different privacy budgets ε.

그림 4는 프라이버시 예산 $\varepsilon \in \{0.5, 1.0, 2.0, 3.0\}$에 따른 세 가지 방법의 유틸리티 변화를 비교한 결과이다. SDP-NS와 SDP-S의 경우에는 서로 다른 마스킹 비율 p에 대한 결과도 함께 제시하였다. 두 데이터셋 모두에서 ε 값이 작아질수록 유틸리티가 감소하는 경향을 확인할 수 있었는데, 이는 DP에서 잘 알려진 프라이버시-유틸리티 절충 관계를 반영한다. 즉, 프라이버시 예산이 줄어들면 지수 메커니즘의 분포가 완만해져 후보 선택이 무작위적으로 이루어지고, 그 결과 원문과 의미적으로 가장 가까운 후보가 선택될 가능성이 낮아진다. SafeDP-Rewrite 역시 이러한 경향을 따르며, DP 기반 접근법의 이론적 속성을 충실히 유지하고 있음을 확인할 수 있다.

또한 단어 단위 변조 기법인 WP는 모든 ε 및 p 설정에서 SDP-NS와 SDP-S에 비해 현저히 낮은 성능을 보였다. 이는 단어 단위 변조 방식은 변형이 누적되면서 문장 전체의 의미와 구조적 일관성을 쉽게 훼손할 수 있음을 확인하였다. 반면 SafeDP-Rewrite는 문장 단위 후보를 생성한 뒤 DP 기반 선택을 적용하기 때문에 의미 보존 측면에서 훨씬 안정적인 성능을 유지하였다. 특히 제안 기법인 SDP-S는 $\varepsilon = 2$ 기준으로 WP 대비 IMDB에서 약 1.7배, MedQuAD에서 약 2.1배(즉, 2배 이상) 향상된 평균 유틸리티를 보였다.

SDP-NS와 SDP-S를 비교한 결과, 모든 ε과 p 조합에서 SDP-S가 SDP-NS보다 일관되게 높은 성능을 보였다. 이는 STEM 과정에서 품질이 낮거나 중복되는 후보를 제거하고, 의미적으로 다양하면서도 유용한 후보만 남기기 때문이다. 그 결과 지수 메커니즘은 동일한 프라이버시 예산 하에서 더 큰 확률을 고품질 후보에 집중시킬 수 있다. 이러한 효과는 특히 마스킹 비율이 높을 때 두드러지는데, 마스킹이 심해질수록 후보 집합에 잡음과 중복이 증가하기 때문이다. STEM은 이러한 상황에서 후보를 효과적으로 정제하여 유틸리티 저하를 완화하는 역할을 한다.

실험 결과를 종합하면, 제안하는 SafeDP-Rewrite는 STEM과 결합할 때 다양한 조건에서도 안정적으로 높은 유틸리티를 달성함을 확인할 수 있었다. 이러한 성능은 데이터셋, 프라이버시 예산, 마스킹 비율 등 여러 환경에서 일관되게 나타나 제안 기법의 견고함을 뒷받침한다. 특히 단어 수준 변조에 기반한 기존 방법과 달리 SafeDP-Rewrite는 문장 단위의 의미를 효과적으로 보존하며, STEM을 통해 중복 후보를 제거하고 다양성을 관리함으로써 결과 텍스트의 품질을 더욱 향상시킨다. 종합적으로, SafeDP-Rewrite는 LLM을 활용한 실용적이고 효과적인 프라이버시 보존 텍스트 변조 기법임을 입증한다.

VI. Conclusions

본 논문에서는 LLM에 대한 블랙박스 접근만으로 동작하는 DP 기반 텍스트 변조 기법인 SafeDP-Rewrite를 제안하였다. SafeDP-Rewrite는 무작위 마스킹과 지수 메커니즘을 결합하여, DP를 만족하는 출력 문장을 최종적으로 생성한다. 또한 후보 집합에서 저품질이거나 중복되는 문장을 제거하는 STEM 기법을 도입하여, 프라이버시를 유지하면서도 유틸리티를 향상시켰다. 실험 결과, SafeDP-Rewrite는 기존의 단어 수준 변조 방식에 비해 전반적으로 안정적이고 높은 성능을 보였으며, 특히 기존 기법 대비 유틸리티가 평균 1.7~2배 향상되어 다양한 데이터 환경에서도 일관된 성능 향상 효과를 확인할 수 있었다. 따라서 SafeDP-Rewrite는 LLM API만으로도 프라이버시를 보장하면서 의미적으로 충실한 텍스트를 생성할 수 있는 실용적이고 기법이다.

본 연구에서는 Gemini-2.0-Flash 모델을 단일 실험 환경으로 사용하였다. 향후에는 이를 확장하여, 다양한 LLM 간 비교와 언어별·도메인별 일반화 성능 검증을 통해 SafeDP-Rewrite의 적용 가능성을 평가할 계획이다.

ACKNOWLEDGEMENT

This research was funded by a 2024 Research Grant from Sangmyung University. (2024-A000-0102)

REFERENCES

[1] J. R. Saura, D. Ribeiro-Soriano, and D. Palacios-Marques. From user-generated data to data-driven innovation: A research agenda to understand user privacy in digital markets. International Journal of Information Management, vol. 60, 2021. DOI: 10.1016/j.ijinfomgt.2021.102331

[2] J. W. Kim, J. H. Lim, S. M. Moon, and B. Jang. Collecting health lifelog data from smartwatch users in a privacy-preserving manner. IEEE Transactions on Consumer Electronics, vol. 65, no. 3, pp. 369-378, 2019. DOI: 10.1109/TCE.2019.2924466

[3] M. Li, J. Liu, and Y. Yang. Automated Identification of Sensitive Financial Data Based on the Topic Analysis. Future Internet, vol. 16, no. 2, 2024. DOI: 10.3390/fi16020055

[4] C. Dwork. Differential privacy. Proceedings of International Colloquium on Automata, Languages, and Programming, pp. 1-12, Venice, Italy, 2006. DOI: 10.1007/11787006_1

[5] O. Feyisetan, B. Balle, T. Drake, and T. Diethe. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. in Proceedings of the International Conference on Web Search and Data Mining, pp.178-186, 2020. DOI: 10.1145/3336191.3371856

[6] Z. Xu, A. Aggarwal, O. Feyisetan, and N. Teissier. A differentially private text perturbation method using regularized Mahalanobis metric. Proceedings of the Second Workshop on Privacy in NLP, pp. 7-17, 2020. DOI: 10.18653/v1/2020.privatenlp-1.2

[7] R. S. Carvalho, T. Vasiloudis, and O. Feyisetan. TEM: High utility metric differential privacy on text. https://arxiv.org/abs/2107.07928, 2021. DOI: 10.48550/arXiv.2107.07928

[8[C. Meehan, K. Mrini, and K. Chaudhuri. Sentence-level privacy for document embeddings. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, pp. 3367-3380, 2022. DOI: 10.18653/v1/2022.acl-long.238

[9] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth, vol. 1, article 9, 2024. DOI: 10.1007/s44336-024-00009-2

[10[H. Zhou, C. Hu, Y. Yuan, Y. Cui, Y. Jin, and C. Chen. Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities. IEEE Communications Surveys & Tutorials, vol. 27, no. 3, pp. 1955-2005, 2025. DOI: 10.1109/COMST.2024.3465447

[11] J. Mattern, B. Weggenmann, and F. Kerschbaum. The Limits of Word Level Differential Privacy. Findings of the Association for Computational Linguistics: NAACL 2022, pp. 867-881, 2022. DOI: 10.18653/v1/2022.findings-naacl.65

[12] S. Utpala, S. Hooker, and P.-Y. Chen. Locally Differentially Private Document Generation Using Zero Shot Prompting. Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 8442-8457, 2023. DOI: 10.18653/v1/2023.findings-emnlp.566

[13] S. Meisenbacher, M. Chevli, J. Vladika, and F. Matthes. DP-MLM: Differentially Private Text Rewriting Using Masked Language Models. Findings of the Association for Computational Linguistics: ACL 2024, pp. 9314-9328, 2024. DOI: 10.18653/v1/2024.findings-acl.554

[14] M. Tong, K. Chen, J. Zhang, Y. Qi, W. Zhang, N. Yu, T. Zhang, and Z. Zhang. InferDPT: Privacy-Preserving Inference for Black-box Large Language Model. https://arxiv.org/abs/2310.12214, 2023. DOI: 10.48550/arXiv.2310.12214

[15] V. Miglani, A. Yang, A. Markosyan, D. Garcia-Olano, and N. Kokhlikyan. Using Captum to Explain Generative Language Models. Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software, pp. 165-173, 2023. DOI: 10.18653/v1/2023.nlposs-1.19

[16] Y. Chang, B. Cao, Y. Wang, J. Chen, and L. Lin. JoPA:Explaining Large Language Model's Generation via Joint Prompt Attribution. https://arxiv.org/abs/2405.20404, 2024. DOI: 10.48550/arXiv.2405.20404

[17] X. Zhou, Y. Lu, R. Ma, T. Gui, Y. Wang, Y. Ding, Y. Zhang, Q. Zhang, and X. Huang. TextObfuscator: Making Pre-trained Language Model a Privacy Protector via Obfuscating Word Representations. Findings of the Association for Computational Linguistics: ACL 2023, pp. 5459-5473, 2023. DOI: 10.18653/v1/2023.findings-acl.337

[18] X. Yue, M. Du, T. Wang, Y. Li, H. Sun, and S. S. M. Chow. Differential Privacy for Text Analytics via Natural Text Sanitization. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 3853-3866, 2021. DOI: 10.18653/v1/2021.findings-acl.337

[19] H. Chen, F. Mo, Y. Wang, C. Chen, J.-Y. Nie, C. Wang, and J. Cui. A Customized Text Sanitization Mechanism with Differential Privacy. Findings of the Association for Computational Linguistics: ACL 2023, pp. 4606-4621, 2023. DOI: 10.18653/v1/2023.findings-acl.355

[20] D. Bollegala, S. Otake, T. Machide, and K. Kawarabayashi. A Metric Differential Privacy Mechanism for Sentence Embeddings. ACM Transactions on Privacy and Security, vol. 28, no. 2, article 20, pp. 1-34, 2025. DOI: 10.1145/3708321

[21] A. B. Abacha and D. D.-Fushman. A Question-Entailment Approach to Question Answering. BMC Bioinformatics, vol. 20, 2019. DOI: 10.1186/s12859-019-3119-4

[22] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 2011.

[23] Sentence Transformers. https://sbert.net accessed on 15 July 2025.

Authors

Jong Wook Kim received the Ph.D. degree in Computer Science Department, Arizona State University, in 2009. He was a Software Engineer with the Query Optimization Group, Teradata, from 2010 to 2013.

Dr. Kim is currently a Professor with the Department of Computer Science at Sangmyung University. His primary research interests include data privacy and query optimization.