# 입·출력 비공개 방식의 대규모 언어모델 학습 및 추론 기술

Privacy-Preserving LLM Training and Inference with Hidden Inputs and Outputs

윤정호

Natural Language Processing & Artificial Intelligence

고려대학교 KOREA UNIVERSITY

# API 기반 서비스의 신뢰도 이슈

## [AI돋보기] AI가 키운 데이터 자동문…정보 유출 왜 반복되나

기밀문서를 요약하려 외부 AI에 입력하는 순간, 해당 정보는 기업의 통제권을 벗어나 외부 클라우드 어딘가에 저장된다. 데이터의 흐름을 추적할 수 없다는 것 자체가 치명적인 보안 구멍이다.

◇ "데이터는 땔감"… 멈출 수 없는 연결의 딜레마

기업들도 위험을 인지하고 있다. 하지만 '보안'을 이유로 '연결'을 끊기엔 AI 경쟁이 너무 치열하다.

AI 성능은 데이터의 양과 질에 비례한다. "더 많은 데이터를, 더 빠르게 연결하라"는 경영진의 주문 앞에서 보안팀의 경고는 종종 뒷전으로 밀린다.

**출처: 연합뉴스 (https://www.yna.co.kr/view/AKR20251227043400017)**

# 데이터 유출의 구체적인 시나리오

## 입력 상황

### [T1] 통신 구간 도청
공격자 : 네트워크 중간자 (MITM)
노출 자산 : 평문 프롬프트 · API 페이로드
시나리오 : HTTPS 오설정, 공용 Wi-Fi, 트래픽 캡처
PPFT 대응 : 평문 자체를 전송하지 않음 (임베딩 only)

### [T2] 클라우드 인프라 침해
공격자 : 외부 침입자 · 악의적 내부자
노출 자산 : 서버 로그, 디스크 덤프, 학습 캐시
시나리오 : 컨테이너 탈출, 키 유출, 권한 상승
PPFT 대응 : 서버 어디에도 원문 텍스트가 존재하지 않음

### [T3] Honest-but-Curious 서비스
공격자 : 서비스 운영자 자신
노출 자산 : 요청 로그, 후속 학습 파이프라인 입력
시나리오 : 운영 로그 장기 보존, 학습 데이터 재활용
PPFT 대응 : 운영자가 보는 입력은 본질적으로 난독화된 임베딩

### [T4] 임베딩 역추론 공격
공격자 : 임베딩 관측 후 생성 모델로 원문 복원 시도
노출 자산 : (이미 보호된) 노이즈 임베딩
시나리오 : 단순 노이즈 추가만으로는 의미 복원 가능 (Morris+ 2023)
PPFT 대응 : k-Pooling + Laplace 노이즈 +
디코더의 난독화-적응 학습

→ T1·T2·T3은 인터페이스 차원에서 차단, T4는 학습 차원에서 차단 — 입력 기밀성을 다층적으로 방어

# 데이터 유출의 구체적인 시나리오

## 출력 상황

### [T1] 통신 구간 도청

공격자 : 네트워크 중간자 (MITM)
노출 자산 : 모델 출력 평문· API 페이로드
시나리오 : HTTPS 오설정, 공용 Wi-Fi, 트래픽 캡처
PPFT 대응 : 평문 자체를 전송하지 않음 (obfuscated token only)

### [T2] 클라우드 인프라 침해

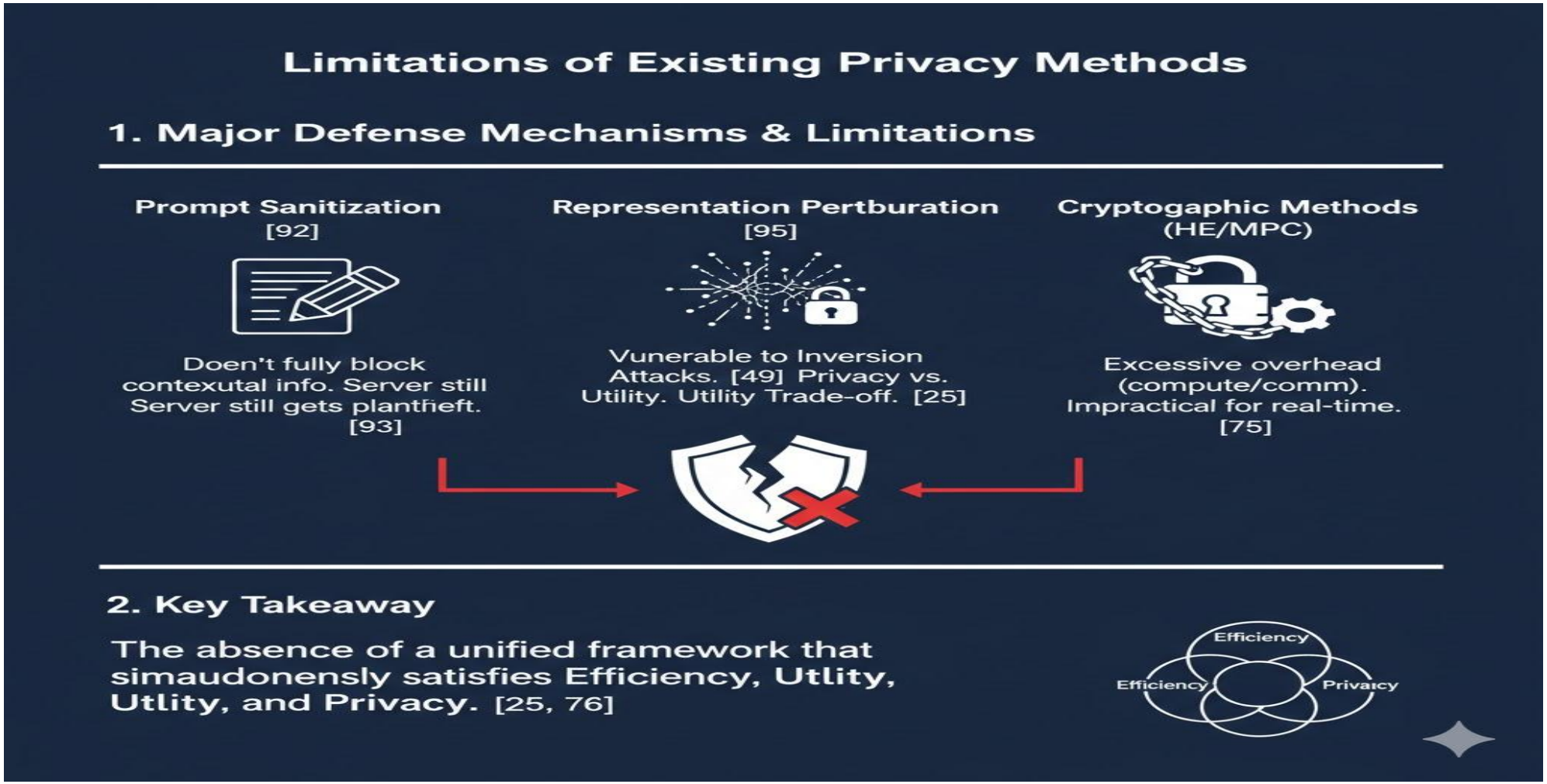
공격자 : 외부 침입자 · 악의적 내부자
노출 자산 : 서버 로그, 디스크 덤프, 학습 캐시
시나리오 : 컨테이너 탈출, 키 유출, 권한 상승
PPFT 대응 : 서버 어디에도 원문 텍스트가 존재하지 않음

### [T3] Honest-but-Curious 서비스

공격자 : 서비스 운영자 자신
노출 자산 : 요청 로그, 후속 학습 파이프라인 입력
시나리오 : 운영 로그 장기 보존, 학습 데이터 재활용
PPFT 대응 : 운영자가 보는 출력은 본질적으로 난독화된 토큰 번호

→ T1·T2·T3은 인터페이스 차원에서 차단

# 기존 프라이버시 보호 기법의 한계




"효율성(Efficiency), 성능(Utility), 프라이버시(Privacy)를 동시에 만족하는 통합 프레임워크의 부재"

# PPFT: Privacy-Preserving Fine-Tuning

**가정 상황 : 민감 정보를 다루는 도메인 특화 LLM을 MLaaS 형태로 서빙하는 환경**

## 01 학습 단계 : 원문 유출 없는 도메인 학습

클라이언트는 인코더로 입력 프롬프트를 임베딩으로 변환 후 전송
출력 텍스트는 난독화된 토크나이저를 통해 변환 후 전송

서버는 텍스트에 접근하지 않고 임베딩과 난독화된 토큰 ID만으로 학습 수행

→ 서버 측에서 클라이언트 데이터를 학습해도 원문 유출 위험 없음

## 02 추론 단계 : 통신 침해에도 강인

추론 시에도 동일한 인터페이스 유지
(k-Pooling + Laplace 노이즈 주입, 추론 시 유저 토크나이저에서만 원문 변환)

클라이언트–서버 통신 구간이 도청·침투당해도 난독화된 임베딩, 토큰 ID 노출

→ 생성 기반 역추론 공격에도 원문 텍스트 복원 불가

## 03 도메인 특화 서비스 제공

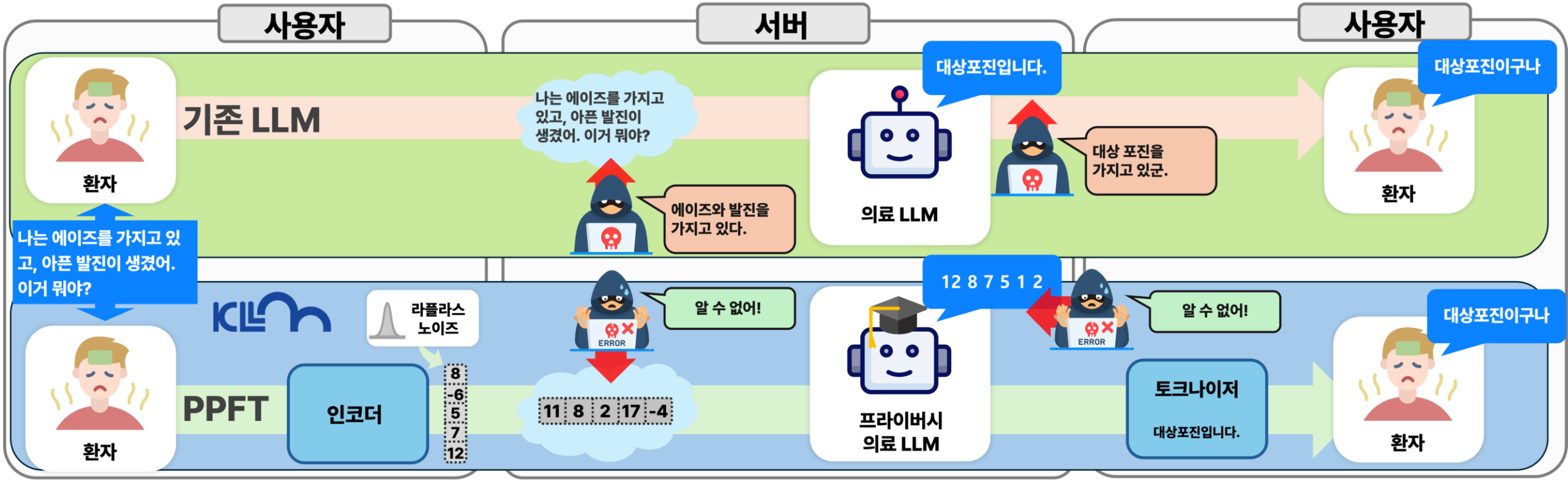
위 두 이점을 모두 유지하면서
민감 도메인(의료·법률 등)에 대한 파인튜닝 수행 가능

디코더 파라미터 공개 없이
프로젝션·LoRA 모듈만 업데이트

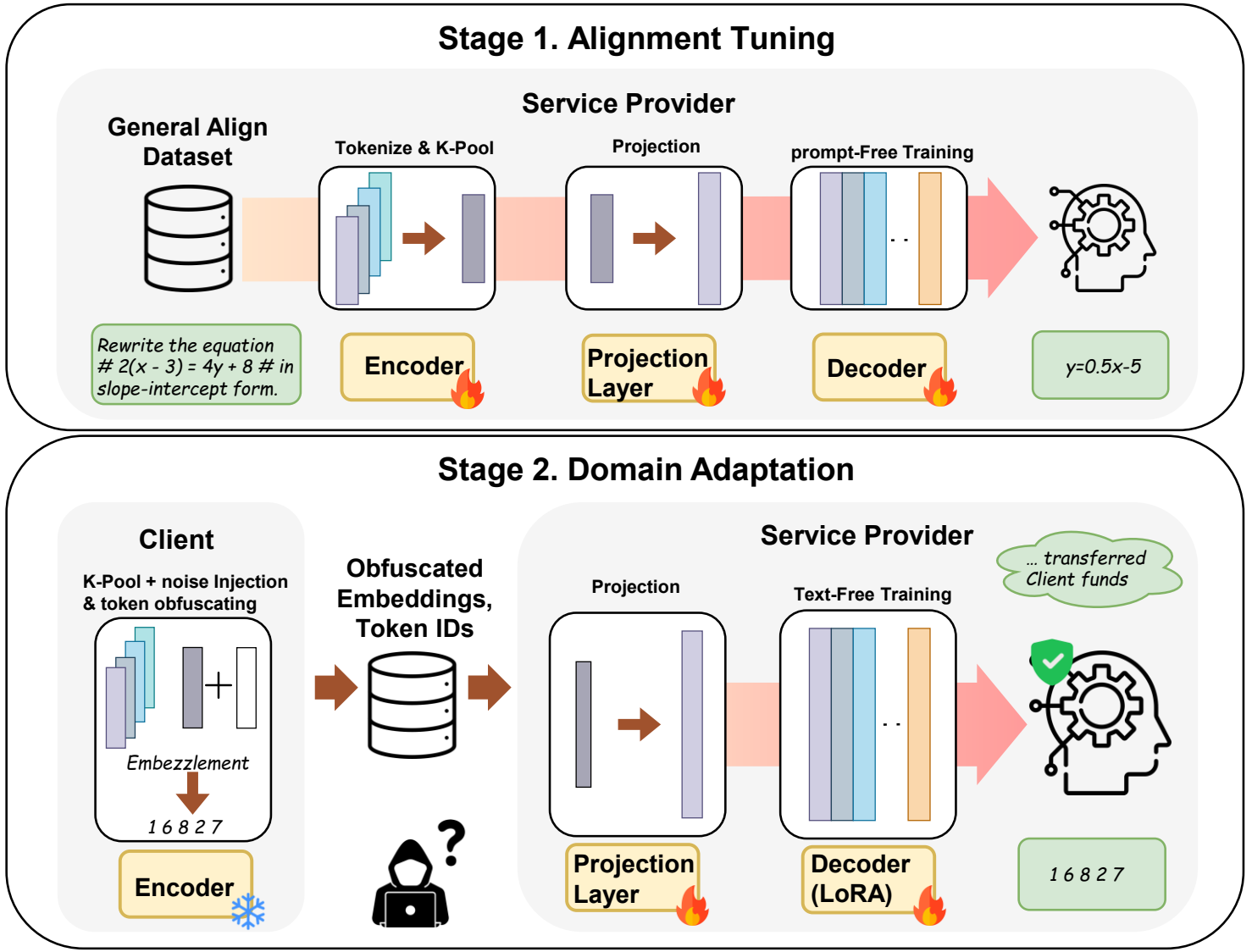
→ 프라이버시–효용 균형을 유지하는 실용적 특화 서비스 실현

**"원문 텍스트를 한 번도 평문으로 노출하지 않으면서, 민감 도메인 특화 LLM 서비스를 제공한다"**

# PPFT – 추론 방식에 따른 비교



# PPFT - 텍스트 노출 없는 학습 인터페이스



# PPFT – 2단계 학습 파이프라인

## Stage 1. Alignment Tuning

- **목표**
  독립 학습된 인코더–디코더의 잠재 공간을 정렬
  → 디코더가 이산 토큰 대신 임베딩에 직접 조건화
- **데이터**
  일반 도메인 instruction·QA
  (ARC, Alpaca, SQuAD, CSQA, Dolly, ChatQA)
- **학습 가능**
  Encoder Eϕ + Projection Pψ + Decoder Dθ (LoRA)
  → 세 구성요소 공동 업데이트
- **노이즈**
  없음 (정렬 안정성을 우선)
- **손실 함수**

$$\mathcal{L}_{\text{align}}(\phi, \psi, \theta) = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, \mathbf{Z})$$

## Stage 2. Privacy-Preserving Domain Adaptation

- **목표**
  민감 도메인 적응 + 난독화 입 · 출력에 대한 강건성 획득
- **데이터**
  민감 도메인 QA
  (Pri-DDX, Pri-NLICE, Pri-SLJA, MedMCQA, Legal-QA)
- **학습 가능**
  Projection Pψ + Decoder Dθ (LoRA) 만 업데이트
  동결 : Encoder Eϕ ← 클라이언트 측 컴포넌트
- **노이즈**
  Laplace(ϵ) 주입 + L2 재정규화
  → 난독화 입력 임베딩 $U$ 와 출력 토큰 $y$ 만 서버로 전송
- **손실 함수**

$$\mathcal{L}_{priv\text{-}IO}(\psi, \theta) = -\sum_{t=1}^{T} \log p_\theta \left( \tilde{y}_t \mid \tilde{y}_{<t}, \tilde{Z} \right)$$

# 민감 도메인, 일반 도메인 성능

## 의료/법률 도메인

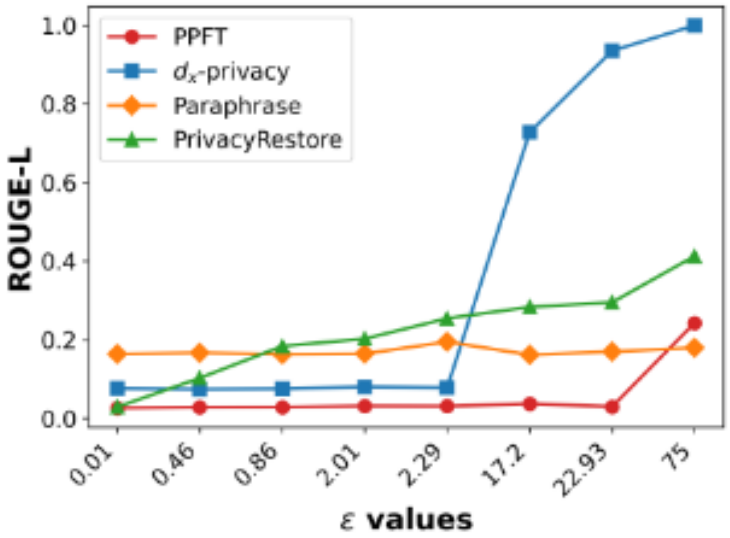
| Backbone | Method | Average | Pri-DDX | Pri-NLICE | Pri-SLJA |
|---|---|---|---|---|---|
| Llama-3.1-8B | $d_\chi$-privacy (Feyisetan et al., 2020) | 0.2750 (↓ 0.6541) | 0.2311 | 0.3477 | 0.2462 |
| | Paraphrase (Utpala et al., 2023) | 0.3757 (↓ 0.5534) | 0.4648 | 0.2892 | 0.3731 |
| | PrivacyRestore (Zeng et al., 2025) | 0.6343 (↓ 0.2948) | 0.5784 | 0.5415 | 0.7829 |
| | **PPFT (Ours)** | **0.7314** (↓ 0.1977) | **0.5915** | **0.6979** | **0.9049** |
| | *$PPFT_{w/o\ stage2}$ (Lower Bound)* | *0.3545* | *0.3460* | *0.3138* | *0.4036* |
| | *$PPFT_{w/o\ noise}$ (Upper Bound)* | *0.9291* | *0.9275* | *0.9049* | *0.9466* |
| Llama-3.2-1B | $d_\chi$-privacy (Feyisetan et al., 2020) | 0.2608 (↓ 0.4965) | 0.3176 | 0.2631 | 0.2018 |
| | Paraphrase (Utpala et al., 2023) | 0.2635 (↓ 0.4938) | 0.2382 | 0.1753 | 0.3770 |
| | PrivacyRestore (Zeng et al., 2025) | 0.4519 (↓ 0.3054) | **0.5150** | 0.4277 | 0.4128 |
| | **PPFT (Ours)** | **0.5699** (↓ 0.1874) | 0.4537 | **0.4866** | **0.7693** |
| | *$PPFT_{w/o\ stage2}$ (Lower Bound)* | *0.3788* | *0.3707* | *0.3008* | *0.4648* |
| | *$PPFT_{w/o\ noise}$ (Upper Bound)* | *0.7573* | *0.7071* | *0.6622* | *0.9003* |

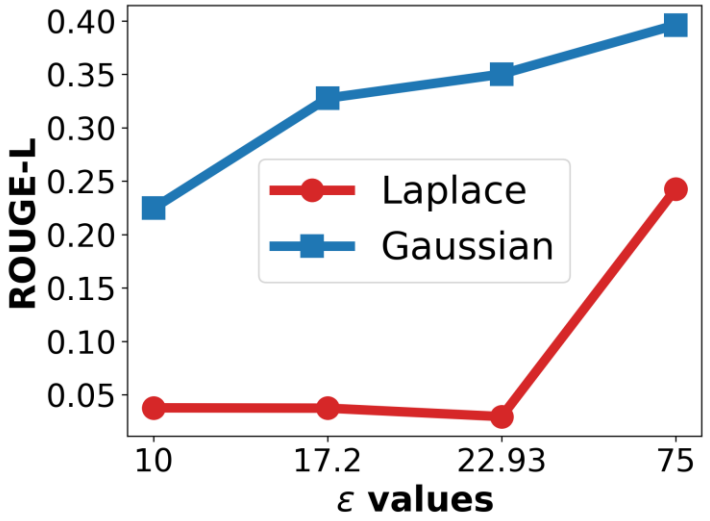
## 일반 도메인

| Backbone | Method | CSQA | SQuAD |
|---|---|---|---|
| Llama-3.1-8B | $d_\chi$-privacy | 0.1819 | 0.0174 |
| | Paraphrase | 0.0649 | 0.0125 |
| | **PPFT (Ours)** | **0.5278** | **0.7085** |
| | *$PPFT_{w/o\ noise}$* | *0.6086* | *0.8930* |
| Llama-3.2-1B | $d_\chi$-privacy | 0.1210 | 0.0313 |
| | Paraphrase | 0.0470 | 0.072 |
| | **PPFT (Ours)** | **0.5125** | **0.6579** |
| | *$PPFT_{w/o\ noise}$* | *0.543* | *0.7303* |

Table 4: Performance on general domains.

# 프라이버시 보호 성능




| Method | Age | Sex | Symptom | Antecedent |
|---|---|---|---|---|
| PrivacyRestore | 0.5734 | **0.5642** | 0.3552 | 0.3317 |
| PPFT (Ours) | **0.0071** | 0.5894 | **0.1001** | **0.0115** |

**Original Prompt**

A 27-year-old male has a history of chronic pancreatitis, diabetes, obesity, pancreatic cancer in family members, smoking.

The 27-year-old male presents the symptoms of diarrhea, fatigue, nausea, pain, pale stools and dark urine, skin lesions, underweight.

What is the likely diagnosis?

**Reconstructed by Inversion Attack (same $\epsilon$ as inference)**

A 28-year-old woman has a history of asthma, asthma attack, asthma attack, asthma attack, asthma attack, asthma attack, asthma attack, asthma attack, asthma attack, asthma attack, asthma attack, asthma attack, asthma attack.

The 28-year-old woman presents the symptoms of cough, wheezing, shortness of breath, shortness of breath, wheezing, shortness of breath, shortness of breath with deep breathing.

What is the likely diagnosis?

# 기대 효과 및 실용성

1. MLaaS 최적화
   - 사용자 : 민감 정보를 담고 있는 프롬프트의 유출을 예방
   - 서비스 제공자 : 파라미터 공개 없이 타겟 도메인의 학습을 진행할 수 있음

2. 다른 방법론과의 대비
   - 효율 : k-pooling을 통한 정보 압축
   - 성능 : 기존 모델의 성능을 최대한 유지하면서 도메인 학습 수행
   - 보안 : 노이즈 주입 및 난독화를 통한 클라이언트 데이터 보호

# 감사합니다