Privacy KULLM

2026. 06. 18.

윤정호

모델의 입력 및 학습 데이터를 난독화해보자.

Towards Privacy-Preserving Large Language Model: Text-free Inference Through Alignment and Adaptation

Jeongho Yoon[1], **Chanhee Park**[1], **Yongchan Chun**[1], **Hyeonseok Moon**[2], **Heuiseok Lim**[1*]
[1]Department of Computer Science and Engineering, Korea University
[2]Samsung Mobile eXperience Business
{aa007878,pch7678,cyc9805,limhseok}@korea.ac.kr
hyns.moon@samsung.com

API 기반 서비스의 신뢰도 이슈

[AI돋보기] AI가 키운 데이터 자동문…정보 유출 왜 반복되나

AI돋보기 + 구독

송고 2025-12-28 06:33

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

[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

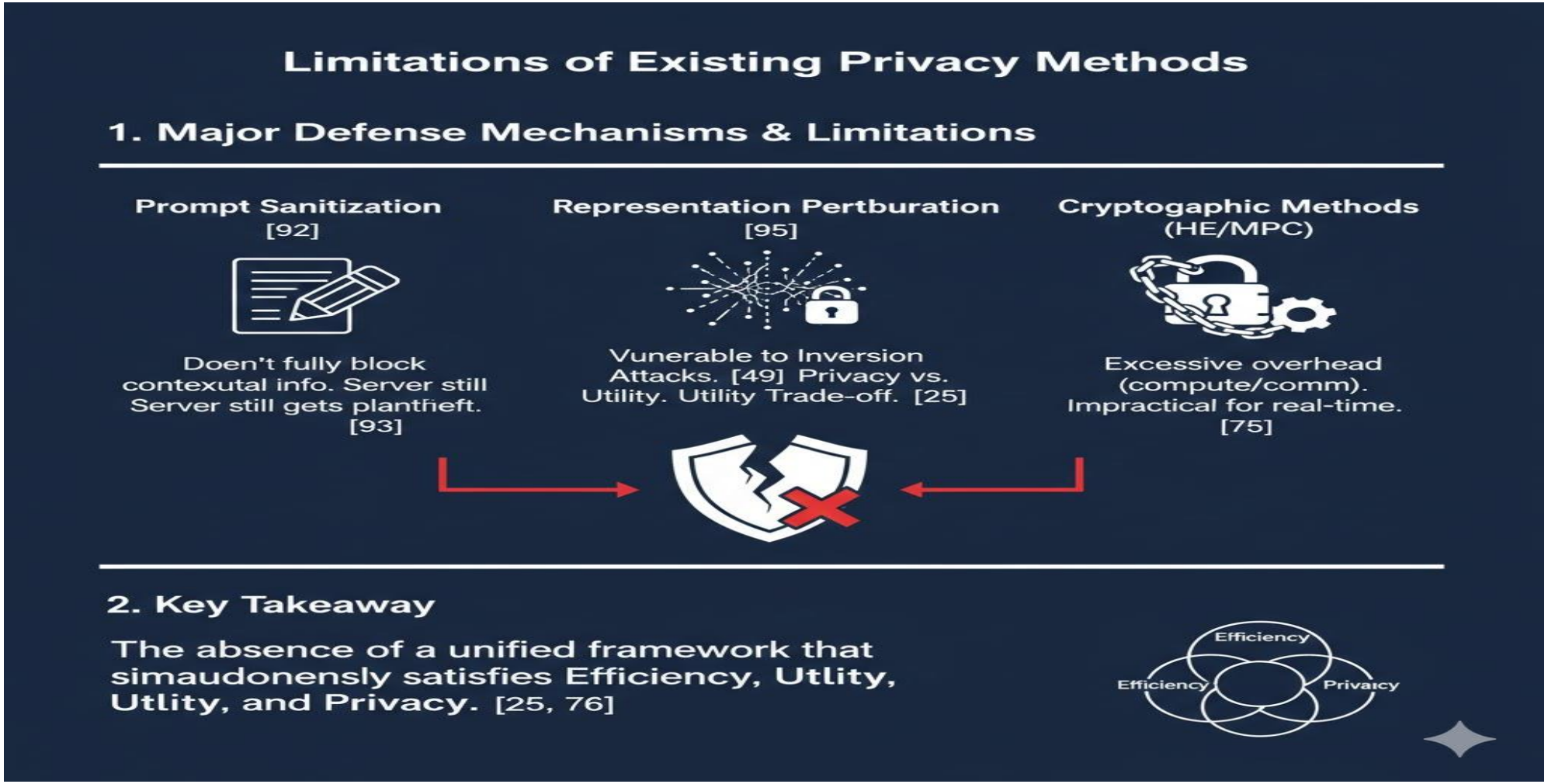
공격자 : 서비스 운영자 자신
노출 자산 : 요청 로그, 후속 학습 파이프라인 입력
시나리오 : 운영 로그 장기 보존, 학습 데이터 재활용
PPFT 대응 : 운영자가 보는 입력은 본질적으로 난독화된 임베딩

[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

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 프롬프트를
임베딩으로 변환 후 전송

서버는 원문 텍스트에 접근하지 않고
임베딩만으로 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)

클라이언트–서버 통신 구간이
도청·침투당해도 난독화된 임베딩만 노출

→ 생성 기반 역추론 공격에도
원문 프롬프트 복원 불가

03 도메인 특화 서비스 제공

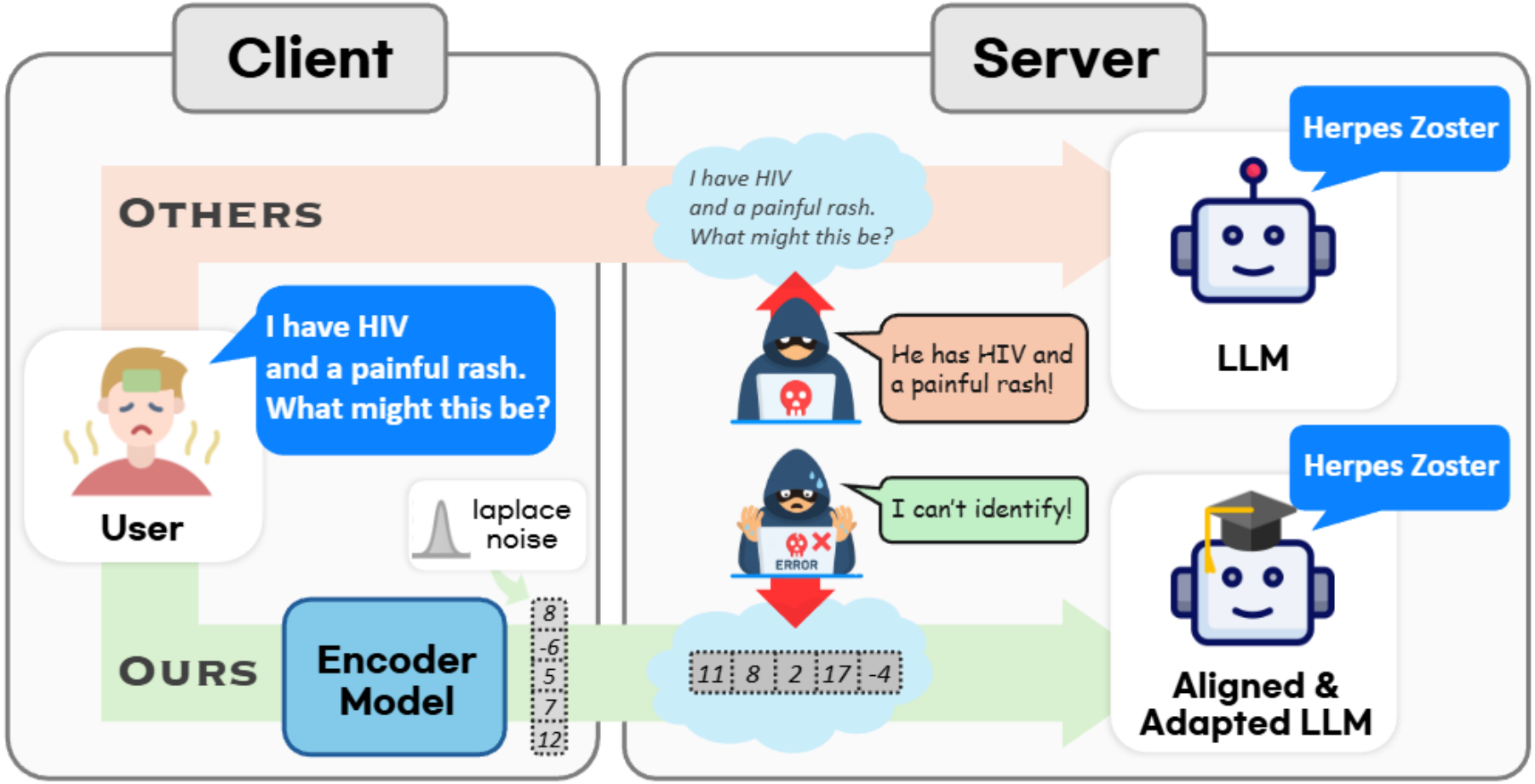
위 두 이점을 모두 유지하면서
민감 도메인(의료·법률 등)에 대한
파인튜닝 수행 가능

디코더 파라미터 공개 없이
프로젝션·LoRA 모듈만 업데이트

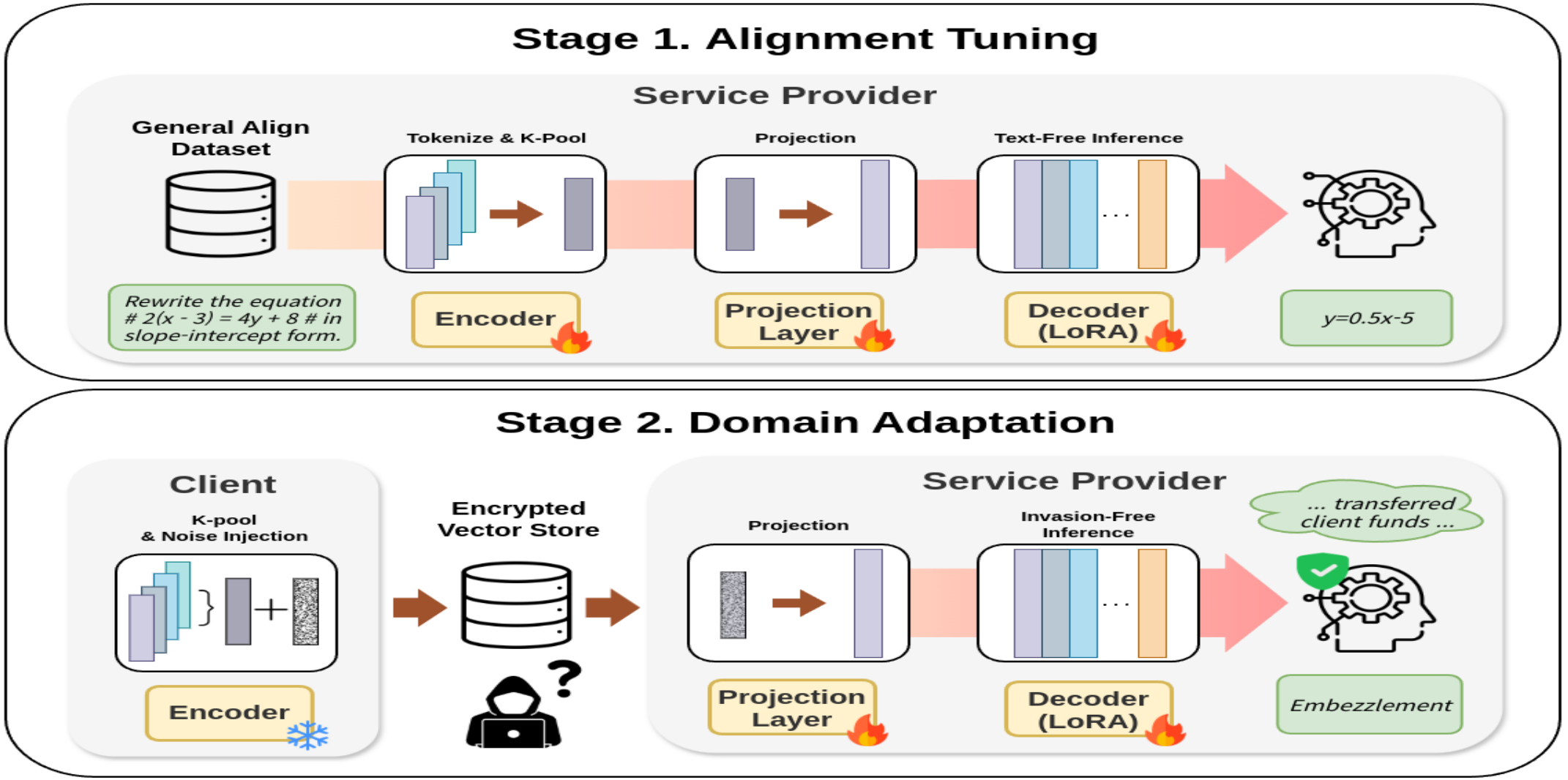
→ 프라이버시–효용 균형을 유지하는
실용적 특화 서비스 실현

"프롬프트를 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 입력에 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
  → 난독화된 임베딩 $\tilde{u}$ 만 서버로 전송
- **손실 함수**

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

민감 도메인, 일반 도메인 성능

의료/법률 도메인

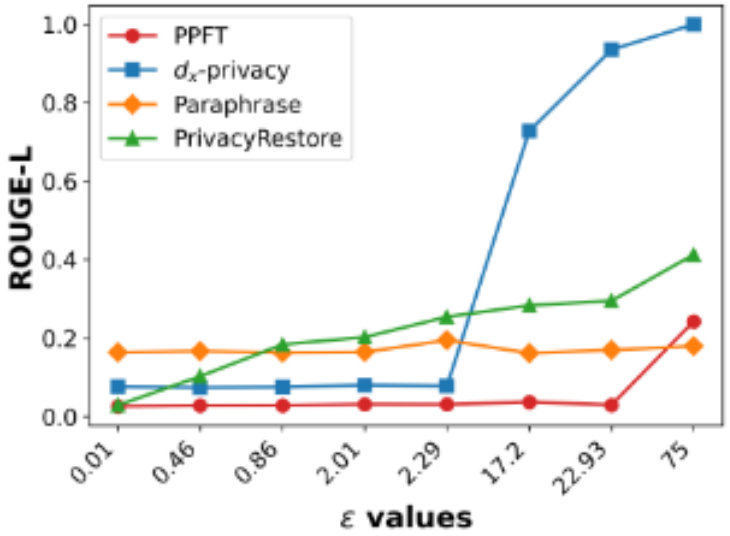
Backbone	Method	Average	Pri-DDX	Pri-NLICE	Pri-SLJA
Llama-3.1-8B	d_χ-privacy (Feyisetan et al., 2020)	0.2750 (↓ 0.6541)	0.2311	0.3477	0.2462
	Paraphrase (Utpala et al., 2023)	0.3757 (↓ 0.5534)	0.4648	0.2892	0.3731
	PrivacyRestore (Zeng et al., 2025)	0.6343 (↓ 0.2948)	0.5784	0.5415	0.7829
	PPFT (Ours)	**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_χ-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*

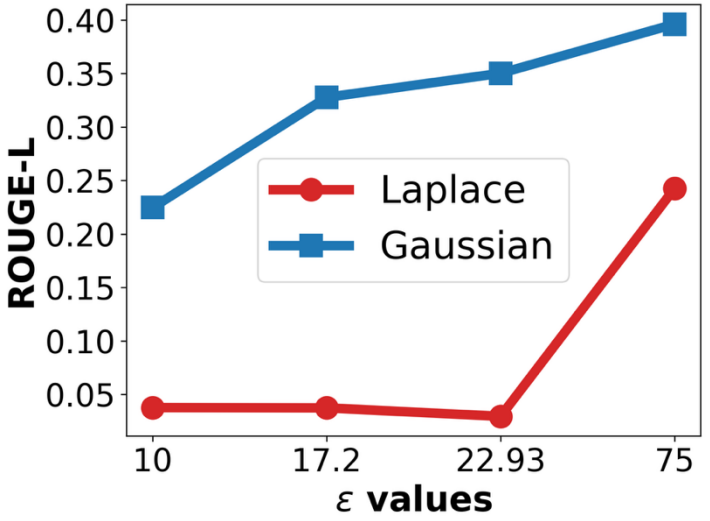
일반 도메인

Backbone	Method	CSQA	SQuAD
Llama-3.1-8B	d_χ-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_χ-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 ϵ 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.

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

모델의 입 출력을 모두 난독화해보자

AlienLM: Alienization of Language for API-Boundary Privacy in Black-Box LLMs

---

Jaehee Kim [1] **Pilsung Kang** [1]

Introduction

1. 모델의 입 출력이 그대로 서버로 전송된다.

2. 데이터 저장 옵션을 끄더라도, 결국 서버에 평문으로 전송되기에 어떻게 사용될지 모른다.

=> 텍스트 수준에서 보안화 작업이 진행되며, 모델의 가중치 접근은 없게, 추론 시점의 프롬프트 및 답변 노출을 줄이는 것이 목표

AlienLM

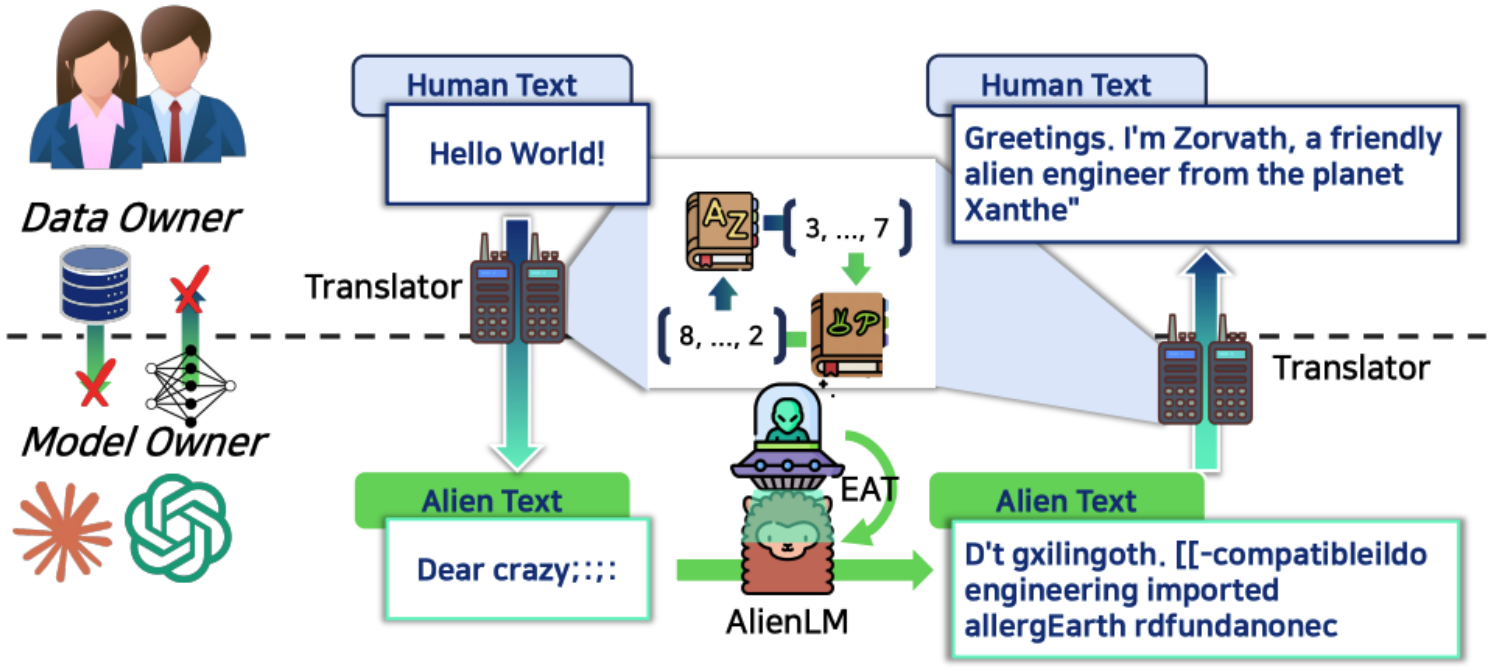


Figure 1. Overview of ***AlienLM***: client-side translation via a vocabulary-level bijection, server-side processing on alien text, and lossless decoding back to plaintext using shared token IDs under a permuted mapping.

natural

Q: Write a Python function to reverse only the vowels of a given string.
A:

```
def reverse_vowels(s):
    vowels = 'aeiouAEIOU'
    s_list = list(s)
    left, right = 0, len(s) - 1
    while left < right:
        if s_list[left] not in vowels:
            left += 1
        elif s_list[right] not in vowels:
            right -= 1
        else:
            s_list[left], s_list[right] = s_list[right],
                s_list[left]
            left += 1
            right -= 1
    return ''.join(s_list)
```

Alien Language

```
Q: Read the/python functionality
stringByAppendingString,
A: getopt(t productList]tyleAttrr
towels "<<UNICODE>>
iNdEx<<UNICODE>>IOExceptionZa"
helps productList (t] right.  left
100.  düzenlenen(t]  211 right
left?  ...
```

AlienLM – 토큰 교환 과정

1. 특수 토큰 (like EOS)을 제외하고 나머지 모든 토큰을 1대 1 대응
   - 공개된 토크나이저와 vocab만 필요하므로 API 호환성 보장
   - 다른 토큰과 매핑하면 가독성 떨어짐
   - 일대일 관계에서 의미적 관계를 보존하면 효율적 fine-tuning 가능
   - 의미적 관계를 보존하며 토큰 예측을 어렵게 하기 위해 후보 토큰 중 임베딩 거리가 가장 먼 토큰으로 선택

2. 클라이언트 단에서 교환 진행
   - 모든 토큰을 교환할지, 특정 민감 토큰 비율을 정해서 교환할지 선택 가능

3. 교환 진행된 데이터셋을 서버에 전송하여 외계어 텍스트를 학습

=>API 제공자를 포함하여 관찰자는 의미 없는 텍스트만 보게 되므로 평문 노출이 대촉 감소

실험 결과

Table 1. Main results across four backbones (accuracy, %; EM for GSM8K). AVERAGE is the unweighted mean; RATIO is the recovery ratio (RR) relative to Oracle. AlienLM uses API-only fine-tuning (AAT) with $\rho = 1$. "-" indicates not run.

Models	Method	MMLU (5-shot)	ARC-Easy (25-shot)	ARC-Challenge (25-shot)	HellaSwag (10-shot)	WinoGrande (5-shot)	TruthfulQA (0-shot)	GSM8K (5-shot)	Average	Ratio
LLaMA 3 8B	Oracle	67.32	84.13	59.39	57.07	74.35	35.25	75.89	64.77	-
	AlienLM	**46.56**	**72.14**	**44.28**	**47.86**	**61.48**	**35.01**	**63.08**	**52.92**	**81.70**
	SentinelLM	29.92	46.34	27.56	38.47	55.09	30.23	31.08	36.96	57.06
	Substitution	25.18	26.39	20.56	26.66	47.59	25.83	1.21	24.77	38.25
	ROT13-ASCII	22.92	25.00	20.05	25.36	49.09	20.93	0.00	23.34	36.03
Qwen 2.5 7B	Oracle	73.50	83.80	57.51	59.66	63.77	47.86	73.09	65.60	-
	AlienLM	**57.87**	**73.11**	**49.23**	**48.43**	**63.69**	**33.78**	**75.21**	**57.33**	**87.40**
	SentinelLM	23.03	35.52	21.16	31.78	49.41	31.95	25.32	31.17	47.51
	Substitution	26.82	29.08	20.22	27.02	50.20	27.78	1.44	26.08	39.76
	ROT13-ASCII	25.51	25.00	20.90	25.03	49.01	21.66	0.00	23.87	36.39
Qwen 2.5 14B	Oracle	78.79	90.36	71.16	71.63	73.72	55.94	72.86	73.49	-
	AlienLM	**65.39**	**79.21**	**53.16**	**50.53**	**66.46**	**38.92**	**80.67**	**62.05**	**84.43**
	SentinelLM	22.95	62.54	42.32	43.38	61.48	34.39	73.09	48.59	66.12
	Substitution	26.56	28.79	18.17	27.15	49.41	28.27	1.52	25.70	34.96
	ROT13-ASCII	26.89	25.67	19.54	25.23	49.72	20.20	0.00	23.89	32.51
Gemma 2 9B	Oracle	71.89	89.35	69.20	60.74	74.59	43.82	74.83	69.20	-
	AlienLM	**54.71**	**75.04**	**48.81**	**50.66**	**60.85**	**35.50**	**70.81**	**56.63**	**81.83**
	SentinelLM	45.88	61.07	41.81	45.38	58.25	33.17	65.73	50.18	72.52
	Substitution	24.51	28.54	19.54	26.37	50.75	25.21	0.30	25.03	36.17
	ROT13-ASCII	23.79	24.58	19.54	25.19	49.41	22.15	0.00	23.52	33.99

학습 세팅
외계어화 비율 1
데이터 : instruction 300k + reasoning 150k

SentinelLM – 모델의 임베딩 스페이스까지 비공개되어 랜덤 1대 1 대응 학습
=> 단순 랜덤은 모델의 적응 성능을 크게 떨어뜨림

Subsitution – 훈련 없이 토큰 스페이스만 변경하여 인퍼런스 (생각보다 높습니다..)
=> 학습이 필요함을 보여줌

ROT13-ASCII – 토큰 단위가 아닌 케릭터 단위로 다 변동한 뒤에 학습 진행
=> 서브워드 경계를 무너뜨려 모델의 성능이 붕괴

실험 결과

Table 3. Recovery success rates across observer scenarios (token recovery for O1/O3, BLEU for O2).

Method	Scenario	Success
Frequency analysis	O1	<0.01%
LLM-based decoding	O2	BLEU <12
MT-based decoding (NLLB)	O2	BLEU <12
Known-plaintext (1K pairs)	O2	<0.22%
Weight-based mapping	O3	<0.11%

O1 – 빈도 분석을 통한 token 맞추기
⇒ **일부 상위 토큰은 빈도가 매우 높아 맞추기 쉽지만 나머지는 분포가 평탄하기에 복구할 수 없음**

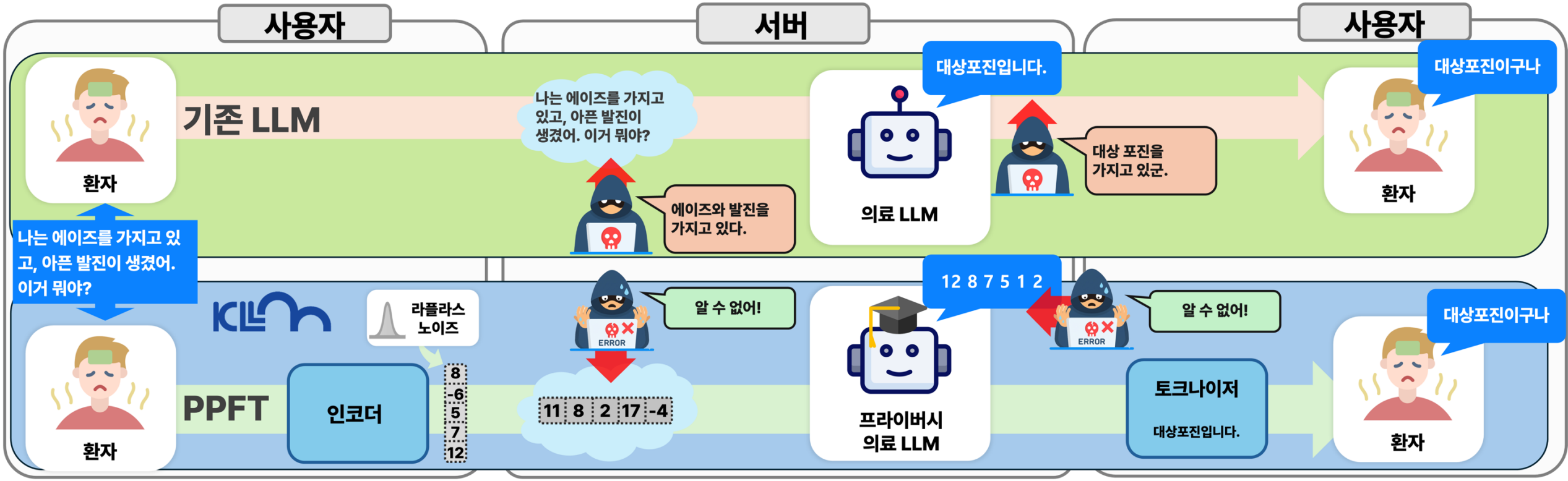
O2 – few-shot 제공하여 번역기로 번역하기
⇒ **GPT-5, 번역 전문 모델, 2만 토큰에 해당하는 문장을 줬어도 실패함**

O3 – 모델이 유출되어 가중치에 접근 가능하여 토큰 간의 유사도를 통해 가장 높은 토큰으로 대치
⇒ **가장 가까운 토큰으로 고른 것이 아니라 성공 확률이 낮음 (top-10으로 늘어나면 대폭 성공확률 올라감)**

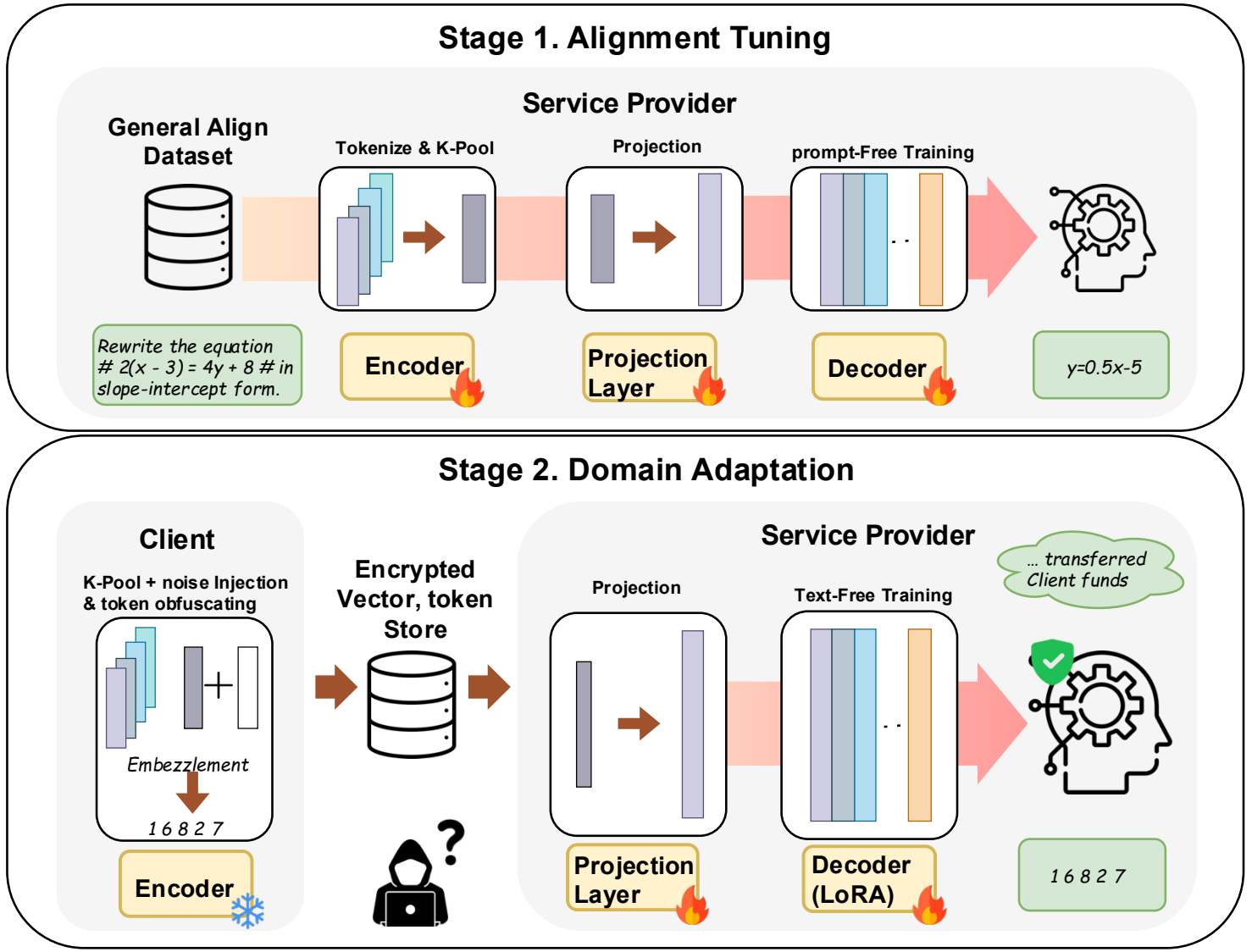
Privacy KULLM

Privacy KULLM

PPFT – 추론 방식에 따른 비교



PPFT - 텍스트 노출 없는 학습 인터페이스



학습 세팅

Encoder – KURE-v1
Decoder – Qwen3.5-2B-Base

Dtype – BF16

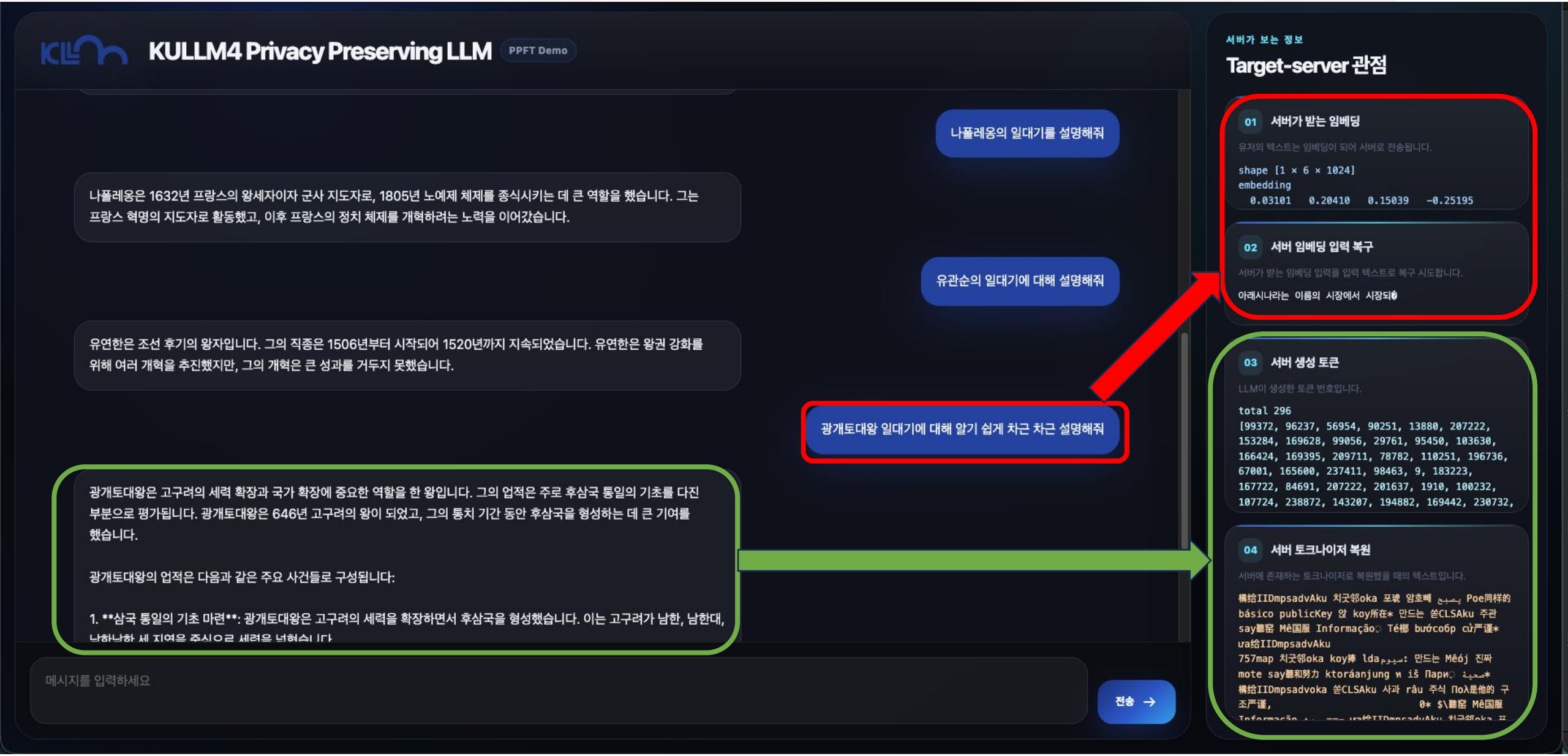
Stage 1

인코더, 디코더 둘 다 full fine-tuning
Pooling size 4
encoder, decoder max length 1024
KULLM 3 DATA + KO, EN Instruction DATA + RAG DATA 해서 총 4M
Per-device batch 8
Gpu: 4
Torch run
Deepspeed Zero-2
Scheduler: cosine

Stage 2

인코더 얼리고, 디코더 LoRA rank64, alpha 128, dropout 0.05
Pooling size 4
encoder, decoder max length 1024
KULLM 3 DATA + KO, EN Instruction DATA + RAG DATA 해서 총 4M
Per-device batch 8
Gpu: 4
Torch run
Deepspeed Zero-2
Scheduler: cosine

Laplace noise
AlienLM과 동일한 방식으로 디코더 토큰 space 섞음


KULLM4 Privacy Preserving LLM
PPFT Demo
나폴레옹의 일대기를 설명해줘
나폴레옹은 1632년 프랑스의 왕세자이자 군사 지도자로, 1805년 노예제 체제를 종식시키는 데 큰 역할을 했습니다. 그는 프랑스 혁명의 지도자로 활동했고, 이후 프랑스의 정치 체제를 개혁하려는 노력을 이어갔습니다.
유관순의 일대기에 대해 설명해줘
유연한은 조선 후기의 왕자입니다. 그의 직종은 1506년부터 시작되어 1520년까지 지속되었습니다. 유연한은 왕권 강화를 위해 여러 개혁을 추진했지만, 그의 개혁은 큰 성과를 거두지 못했습니다.
광개토대왕 일대기에 대해 알기 쉽게 차근 차근 설명해줘
광개토대왕은 고구려의 세력 확장과 국가 확장에 중요한 역할을 한 왕입니다. 그의 업적은 주로 후삼국 통일의 기초를 다진 부분으로 평가됩니다. 광개토대왕은 646년 고구려의 왕이 되었고, 그의 통치 기간 동안 후삼국을 형성하는 데 큰 기여를 했습니다.
광개토대왕의 업적은 다음과 같은 주요 사건들로 구성됩니다:
1. **삼국 통일의 기초 마련**: 광개토대왕은 고구려의 세력을 확장하면서 후삼국을 형성했습니다. 이는 고구려가 남한, 남한대,
메시지를 입력하세요
전송
서버가 보는 정보
Target-server 관점
01 서버가 받는 임베딩
유저의 텍스트는 임베딩이 되어 서버로 전송됩니다.
shape [1 × 6 × 1024]
embedding
0.03101 0.20410 0.15039 -0.25195
02 서버 임베딩 입력 복구
서버가 받는 임베딩 입력을 입력 텍스트로 복구 시도합니다.
아래시나라는 이름의 시장에서 시장되
03 서버 생성 토큰
LLM이 생성한 토큰 번호입니다.
total 296
[99372, 96237, 56954, 90251, 13880, 207222,
153284, 169628, 99056, 29761, 95450, 103630,
166424, 169395, 209711, 78782, 110251, 196736,
67001, 165600, 237411, 98463, 9, 183223,
167722, 84691, 207222, 201637, 1910, 100232,
107724, 238872, 143207, 194882, 169442, 230732,
04 서버 토크나이저 복원
서버에 존재하는 토크나이저로 복원했을 때의 텍스트입니다.
構给IIDmpsadvAku 치긋邻oka 포함 암호빼 يصبح Poe同样的
básico publicKey 않 koy所在* 만드는 쏟CLSAku 주관
say聽窑 Mê国服 Informação Té榔 bướcoбр cử严谨*
ura给IIDmpsadvAku
757map 치긋邻oka koy捧 ldaسبوم: 만드는 Mêój 진짜
mote say聽和努力 ktoráanjung и iš Парк صحية*
構给IIDmpsadvoka 쏟CLSAku 사과 râu 주식 Пολ是他的 구
조严谨, 0* $\聽窑 Mê国服

KULLM4 Privacy Preserving LLM

PPFT Demo

오늘 야식으로 치킨 먹고 싶은데 치킨 브랜드랑 추천하는 브랜드 설명해줘

고양이는 다양한 종류가 있으며, 각 종류마다 맛과 식감이 다릅니다. 다음은 몇 가지 인기 있는 치킨 종류와 추천 브랜드입니다.

1. **고등조개 치킨**:
   - **장점**: 고등조개 치킨은 고등조개의 독특한 맛이 특징입니다. 고등조개의 단맛과 고기를 함께 먹으면 맛의 균형을 맞출 수 있습니다.
   - **bstivy 브랜드**:
     - **고등조개 치킨**: 고등조개의 맛을 잘 살려주며, 다양한 조리법을 통해 맛을 조절할 수 있습니다.
     - **고등조개 치킨 스테이크**: 고등조개 스테이크는 고등조개의 맛을 잘 살려주며, 스테이크 형태로 조리해 맛을 더욱 돋보이게 합니다.

2. **고등조개 치킨 스테이크**:
   - **장점**: 고등조개 스테이크는 고등조개의 맛을 잘 살려주며, 스테이크 형태로 조리해 맛을 더욱 돋보이게 합니다.
   - **bstivy 브랜드**:
     - **고등조개 스테이크**: 고등조개의 맛을 잘 살려주며, 스테이크 형태로 조리해 맛을 더욱 돋보이게 합니다.
     - **고등조개 스테이크**: 고등조개의 맛을 잘 살려주며, 스테이크 형태로 조리해 맛을 더욱 돋보이게 합니다.

3. **고등조개 치킨 샐러드**:
   - **장점**: 고등조개 치킨 샐러드는 고등조개의 맛을 잘 살려주며, 샐러드 형태로 조리해 맛을 더욱 돋보이게 합니다.
   - **bstivy 브랜드**:
     - **고등조개 치킨 샐러드**: 고등조개의 맛을 잘 살려주며, 샐러드 형태로 조리해 맛을 더욱 돋보이게 합니다.
     - **고등조개 치킨 샐러드**: 고등조개의 맛을 잘 살려주며, 샐러드 형태로 조리해 맛을 더욱 돋보이게 합니다.

메시지를 입력하세요

전송 →

서버가 보는 정보

Target-server 관점

01 서버가 받는 임베딩

유저의 텍스트는 임베딩이 되어 서버로 전송됩니다.

```
shape [1 × 6 × 1024]
```

02 서버 임베딩 입력 복구

서버가 받는 임베딩 입력을 입력 텍스트로 복구 시도합니다.

03 서버 생성 토큰

LLM이 생성한 토큰 번호입니다.

```
total 510
[152182, 217222, 119913, 180005, 220236,
101412, 195008, 25, 187362, 220236, 148967,
```

04 서버 토크나이저 복원

서버에 존재하는 토크나이저로 복원했을 때의 텍스트입니다.

```
ópcik磋商 mịn päät稚 일단: 링 päätاض 간빼 Ukr尔捧
Ténшд* grátis 캠 الأم παί ballot 고头 päät肘 أعلن
çer所在*                                    0* $\óp扫Lö给 고
头==-


( $\욱蚁==- 치扫Lö给 고头Aku 치扫Lö를变 مسرح أغ 所
在* 치扫Lö를变 Kø统빼 치的主导 çık ॐ你那 간oka
ঙhanno 총塘 Lam 캐나다*


( $\尊udu çer==-


( $\óp扫Lö给 고头==- 치扫Lö를变 ?!pet kid 외郭창俾:
```

KULLM4 Privacy Preserving LLM

PPFT Demo

안녕하세요! 질문이 무엇인지 좀 더 구체적으로 말씀해 주실 수 있나요? 도움이 될 수 있도록 도와드리겠습니다.

넌 누구야

저는 고려대학교에서 만든 AI 챗봇 "KULLM"입니다. 사용자와의 대화를 통해 도움을 주고자 합니다. 어떤 질문이나 도움이 필요하신가요?

세종대왕 맥북던짐사건에 대해 설명해줘

세종대왕은 1453년 고종이 사망한 후 왕위를 장악한 왕입니다. 세종대왕은 왕권 강화를 위해 여러 개혁을 추진했지만, 그 중에서도 특히 명량해방과 함께 만세대왕의 사임과 함께 많은 정치적 혼란이 발생했습니다.

1456년, 세종대왕은 만세대왕의 사임과 함께 왕권 강화를 위해 여러 개혁을 추진했지만, 그 중에서도 특히 명량해방과 함께 만세대왕의 사임과 함께 많은 정치적 혼란이 발생했습니다.

이순신이 누구야 ???

이순신은 조선 후기의 왕자입니다. 1418년 조선의 세조가 사망한 후, 그의 아들 이순신이 왕위에 올랐습니다.

메시지를 입력하세요

전송 →

서버가 보는 정보

Target-server 관점

01 서버가 받는 임베딩

유저의 텍스트는 임베딩이 되어 서버로 전송됩니다.

```
shape [1 × 3 × 1024]
embedding
  -0.04883   0.00510   0.23145   -0.16016
-0.33203   -0.20508   0.20703   0.20410
  -0.19336   -0.08496   0.15039   0.01160
-0.42383   -0.01929   0.46484   -0.07617
```

02 서버 임베딩 입력 복구

서버가 받는 임베딩 입력을 입력 텍스트로 복구 시도합니다.

```
The following is a list of the most common and
common types of mushrooms.

Mushrooms

The following are the most common types of
mushrooms.
```

03 서버 생성 토큰

LLM이 생성한 토큰 번호입니다.

```
total 39
[101244, 95257, 97251, 207222, 230119, 1910,
136538, 237411, 169502, 98463, 9, 38177, 15,
20, 15, 24, 2117, 230119, 29761, 95450, 205782,
101412, 208766, 97968, 1910, 25, 183223,
```

04 서버 토크나이저 복원

서버에 존재하는 토크나이저로 복원했을 때의 텍스트입니다.

```
捧Leo损Aku médicament say高興 koy늘所在*
0509map médicamentoka 포Lõ稚 ضای凉 say: 만드는 ط
```

KULLM4 Privacy Preserving LLM

PPFT Demo

고려대학교에 대해 설명해줘

고려대학교는 1905년에 설립된 대학으로, 서울특별시 고려에 위치한 대학입니다. 고려대학교는 한국의 주요 대학 중 하나로, 다양한 학문 분야에서 뛰어난 성과를 거두고 있습니다. 고려대학교는 학문적 연구뿐만 아니라, 학생들의 학습과 성장을 돕기 위해 다양한 프로그램과 서비스를 제공합니다.

고려대학교는 다양한 학문 분야의 교수진과 학생들을 보유하고 있으며, 특히 과학, 기술, 사회, 문화 등 여러 분야에서 뛰어난 성과를 거두고 있습니다. 또한, 고려대학교는 학생들에게 다양한 학습 기회를 제공하고 있으며, 연구 프로젝트, 워크숍, 세미나 등 다양한 학습 활동을 통해 학생들의 학습 경험을 풍부하게 하고 있습니다.

고려대학교는 또한 대학 내의 다양한 단체와 네트워크를 통해 사회적 책임과 사회적 기여를 강조하고 있으며, 다양한 사회적 활동과 연구 프로젝트를 추진하고 있습니다.

10 * 10 + 3 * 6 =

이 식을 풀려면 곱셈과 나눗셈을 수행한 다음 합계를 수행해야 합니다.

먼저 10에 10을 곱합니다:
10 * 10 = 100

다음으로 3에 6을 곱합니다:
3 * 6 = 18

메시지를 입력하세요

전송 →

서버가 보는 정보

Target-server 관점

01 서버가 받는 임베딩

유저의 텍스트는 임베딩이 되어 서버로 전송됩니다.

```
shape [1 × 3 × 1024]
embedding
  0.23340   −0.16113   0.23926   0.45313
```

02 서버 임베딩 입력 복구

서버가 받는 임베딩 입력을 입력 텍스트로 복구 시도합니다.

```
The following table shows the number of people
who have been registered to vote in the last
```

03 서버 생성 토큰

LLM이 생성한 토큰 번호입니다.

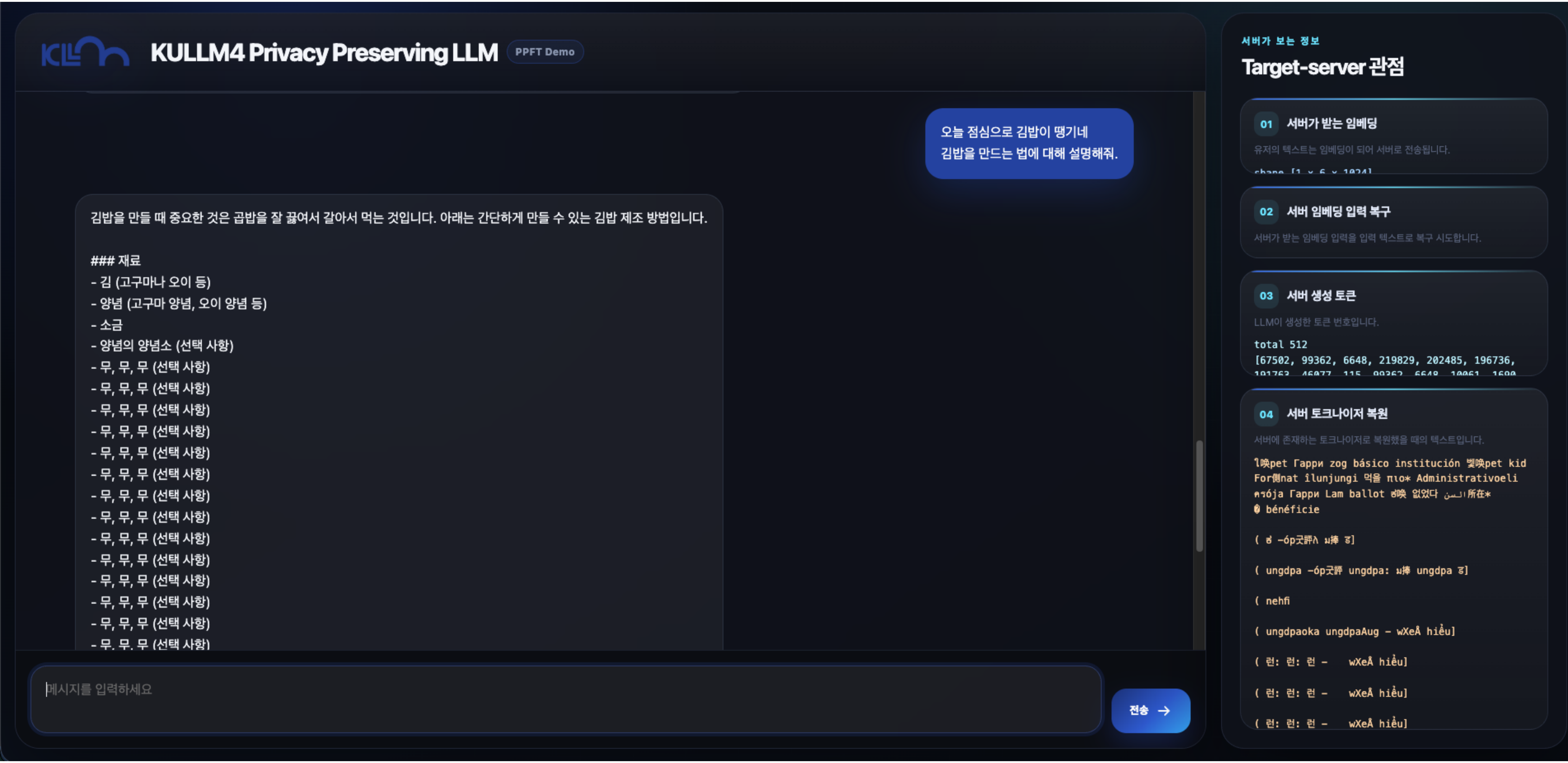
```
total 107
[101244, 13889, 6648, 223692, 119574, 46077,
115, 96542, 169395, 78354, 16228, 157, 96542,
6648, 58648, 97968, 171384, 191995, 127867,
58648, 111788, 187479, 9, 3388, 100310, 69468,
38177, 15, 16, 96275, 38177, 15, 16, 6648,
```

04 서버 토크나이저 복원

서버에 존재하는 토크나이저로 복원했을 때의 텍스트입니다.

```
捧 Ukrpet Fid十七条 墜材畔 Ainnh材pet bells凉 متابع
sē鳄鱼 bells雷达 ات*                                笼
hra
01集
01pet 墜科创板,

01 /
01 +
011                                  逆的g6p
```


KULLM4 Privacy Preserving LLM
PPFT Demo
오늘 점심으로 김밥이 땡기네
김밥을 만드는 법에 대해 설명해줘.
김밥을 만들 때 중요한 것은 곱밥을 잘 끓여서 갈아서 먹는 것입니다. 아래는 간단하게 만들 수 있는 김밥 제조 방법입니다.
재료
- 김 (고구마나 오이 등)
- 양념 (고구마 양념, 오이 양념 등)
- 소금
- 양념의 양념소 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
- 무, 무, 무 (선택 사항)
메시지를 입력하세요
전송
서버가 보는 정보
Target-server 관점
01 서버가 받는 임베딩
유저의 텍스트는 임베딩이 되어 서버로 전송됩니다.
shape [1 × 6 × 1024]
02 서버 임베딩 입력 복구
서버가 받는 임베딩 입력을 입력 텍스트로 복구 시도합니다.
03 서버 생성 토큰
LLM이 생성한 토큰 번호입니다.
total 512
[67502, 99362, 6648, 219829, 202485, 196736,
04 서버 토크나이저 복원
서버에 존재하는 토크나이저로 복원했을 때의 텍스트입니다.
pet Гаppи zog básico institución pet kid
For nat ilunjungi 먹을 πιο* Administrativoeli
ója Гаppи Lam ballot 없었다 所在*
bénéficie
(ungdpa -óp ungdpa: ungdpa]
(nehfi
(ungdpaoka ungdpaAug – wXeÅ hiểu]
(런: 런: 런 – wXeÅ hiểu]
(런: 런: 런 – wXeÅ hiểu]
(런: 런: 런 – wXeÅ hiểu]

감사합니다