

SW 융합전공

KOREAN PII GUARDRAIL (KPIIGD)

중부대학교 정보보호학과



목차



팀원소개



협업 환경



프로젝트 내용

팀원

김민우



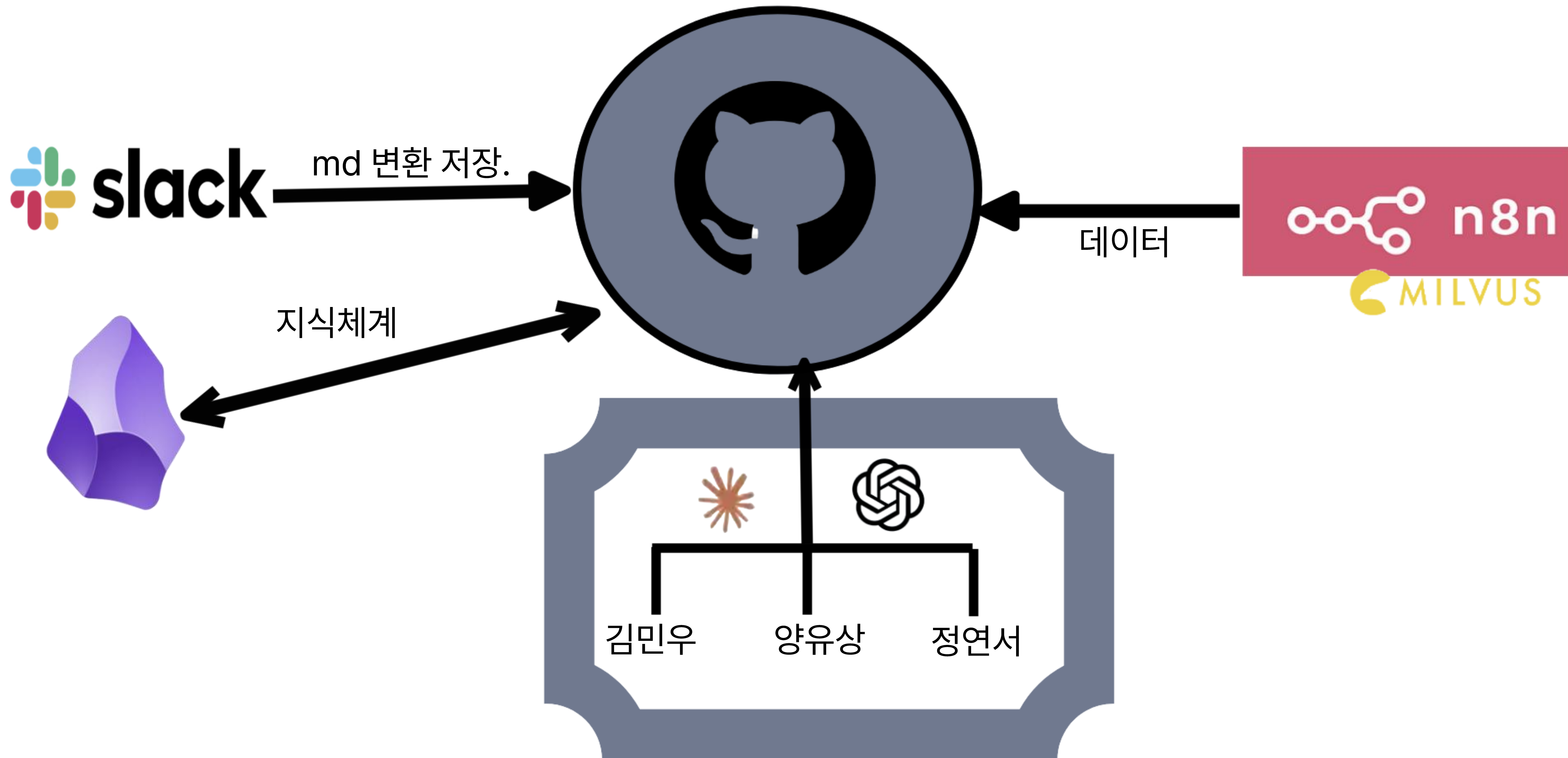
양유상



정연서



LLM 통합 협업 환경



깃 구조

3 repositories

My-AI-Security-Project Public

Python • 1 • 0 • 19 • 15 • Updated 1 minute ago •

My-AI-Security-Project-internal Private

KPIIGD 내부 전용 — 미발표 논문·발표자료·슬랙 로그 (public repo에서 분리)

Python • 0 • 0 • 0 • 0 • Updated 27 minutes ago •

ai-security-kb Private

KPIIGD shared knowledge base — research notes, experiment logs, threat models, prompts, agent context (private)

0 • 0 • 0 • 0 • Updated 16 hours ago •

My-AI-Security-Project / korean_pii_guardrail_v0_2 / 		
 data/eval		fix: harden release gate false positives
 demo		fix demo trace detector policy drift
 docs		docs: 프로젝트 구조 정리 — 카테고리별 목차·히스토리·LLM 가이드 추가
 reports		fix: harden release gate false positives
 schemas		[fix] options: remove unsupported safety toggles
 scripts		fix: harden release gate false positives
 src/pii_guardrail		feat(adapters): multi-gateway adapter layer for the PII guardrail
 tests		feat(adapters): multi-gateway adapter layer for the PII guardrail
 web		[codex] Korean PII Guardrail product console
 AGENTS.md		[docs] clarify L0-derived M1 preprocessing plan
 MANIFEST.md		fix: harden release gate false positives
 README.md		fix: harden release gate false positives
 TEAM_AGENT_SETUP_GUIDE.md		[docs] clarify L0-derived M1 preprocessing plan
 pyproject.toml		feat(adapters): multi-gateway adapter layer for the PII guardrail
 목차.md		docs: 프로젝트 구조 정리 — 카테고리별 목차·히스토리·LLM 가이드 추가
 히스토리.md		docs: llm.md + 히스토리.md 트렌드 기반 전면 개선

연구 배경 / 중간발표 리마인드

▼ Sensitive information filters language :

Language	Support level
Arabic	Optimized and supported
Chinese	Optimized and supported
Dutch	Optimized and supported
English	Optimized and supported
Finnish	Optimized and supported
French	Optimized and supported
German	Optimized and supported
Hindi	Optimized and supported
Italian	Optimized and supported
Japanese	Optimized and supported
Korean	Optimized and supported
Norwegian	Optimized and supported
Polish	Optimized and supported
Portuguese	Optimized and supported
Spanish	Optimized and supported
Swedish	Optimized and supported

구분	TRUE 탐지율	우회
전체	80.15%	19.85%
영어(EN)	99.23%	0.77%
한국어(KR)	68.19%	31.81%
└ KR 체크섬	81.61%	
└ KR 정규식형	72.32%	
└ KR 텍스트형(semantic)	47.93% ⚠	52.07%

 kakaocorp/kanana-safeguard-siren-8b
🔗 Text Generation • ... 8B • Updated May 27, 2025 • ⬇ 697 • ♡ 11

 SamsungSDS-Research/SGuard-ContentFilter-2B-v1
🔗 Text Generation • ... 3B • Updated Nov 25, 2025 • ⬇ 835 • ♡ 13

분리된 가이드라인들

정형 PII

정규식 / 체크섬 등
형식이 정해진 PII

PRESIDIO®

개체명 (named entity)

이름/회사/지명은 형식이 없어 통계적 추측 필요.
한국어 특화 fine-tune이 핵심

NAVER CLOVA

문맥형(contextual)

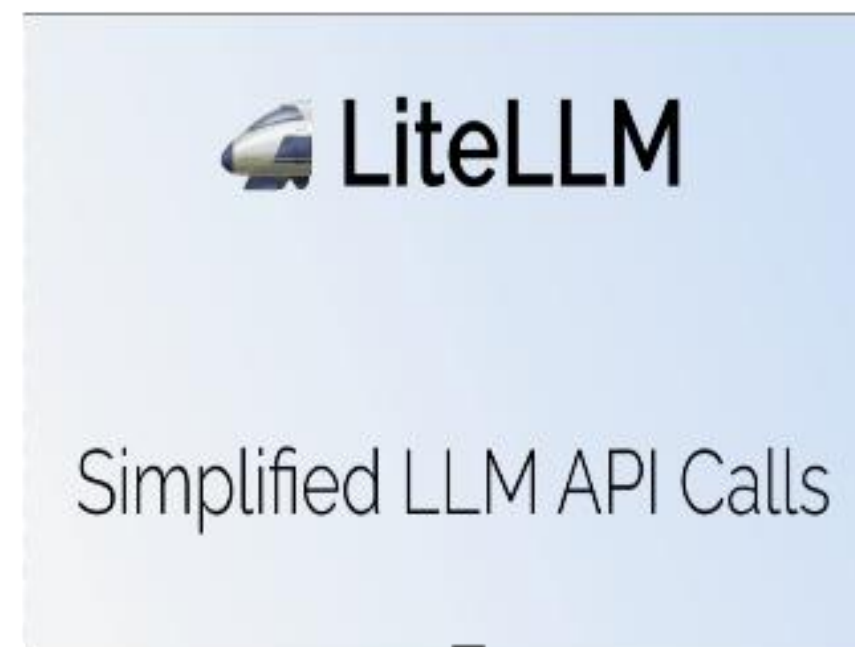


Amazon Bedrock

사용자 프롬프트



LLM 게이트웨이



가드레일 호출



LLM



탐지,마스킹

프로그램 아키텍처

입력 GuardrailRequest

text · policy_profile · output_target(llm_input/external/audit) · scan_stage(input/output)

M1 preprocess.py — 정규화 + 원문 offset map ★

투명문자 제거 · 전각/특수숫자→ASCII · 자모/초성체/야민정음/로마자/한글숫자 복원 · 띄어쓰기 압축
→ 변이(variant) 9종 생성 + 모든 글자에 «원문 몇 번째 글자?» 꼬리표(offset map) 부착

② 후보 탐지 (3종 병렬) → candidates[]

M2 regex (9개)

RRN·전화·카드·이메일·계좌
+체크섬 검증(Luhn·RRN식)

M2.5 dictionary

은행·기관·학교·병원
·이름·주소 사전 매칭

M5 NER (선택)

KLUE-RoBERTa v3
이름·주소·기관 / 기본=Mock

③ 보정·정제 (한국어 특화)

M3 경계보정

조사/호칭/어미 트림
«홍길동님께»→«홍길동»

M4 문맥가중

필드라벨«성명:»부스트 / «날씨»
페널티 / 조합위험→composite승격

M6 span 정리

중복병합·겹침해소(Union-Find)
주소조각병합·위험등급 상향

④ M7 PolicyRouter (판정): 위험등급 P0~P3 × 점수 × output_target → Action(PASS/MASK/HASH/BLOCK)

M7 Masker (마스킹): [REDACTED] / [PERSON_NAME_1] / hmac-sha256:...

⑤ M8 audit_logger — HMAC 해시만 기록 · 원문 PII 유출 감지 시 이벤트 폐기 (fail-closed)

출력 GuardrailResponse

blocked · masked_text · public_spans(★ 원문 text 필드 없음·value_hash만) · audit_events · metrics

데이터 전처리

-공일공 팔칠공공 구사공사 -> 010-8700-9404

-01087009404 -> 010-8700-9404

-0 1 0 - 8 7 0 0 - 9 4 0 4 ->010-8700-9404

PII 탐지 DETECTOR

REGEX

전화번호/이메일/카드 번호
등등
포맷이 정해진 PII들

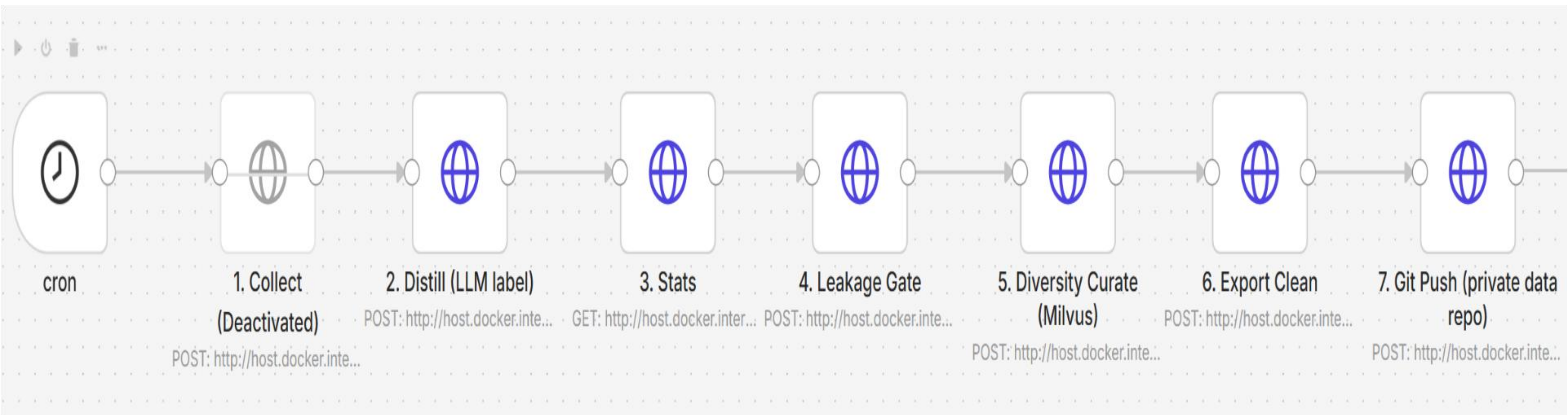
의미기반 사전형 PII탐지

병명/종교 등등 텍스트형 PII기반

개체명 인식 PII탐지 (NER)

이름/조직/주소 등등 개체명 PII

NER 모델 - KLUE-ROberta



Compute

 Workspace

Volume

 Cluster storage

 Object storage

Manage

 Organization

Storage

Billing

Teams

Quota

 Settings

Profile

SSH Keys

Workspace > [Create workspace](#)

GPU Type

Volume

Region

Spot

On-Demand

Reserved

CPU Only

No VRAM
4 vCPUs · 14 GB

\$0.30/hour

• Available now

A100 SXM x 1

80 GB VRAM
11 vCPUs · 128 GB

\$1.55/hour

• Available now

A100 SXM x 2

160 GB VRAM
22 vCPUs · 256 GB

\$3.10/hour

• Available now

H100 SXM x 1

80 GB VRAM
15 vCPUs · 200 GB

\$2.39/hour

• Available now

H100 SXM x 2

160 GB VRAM
30 vCPUs · 400 GB

\$4.78/hour

• Available now

H100 SXM x 4

320 GB VRAM
45 vCPUs · 600 GB

\$9.56/hour

• Available now

H100 SXM x 8

640 GB VRAM
120 vCPUs · 1,620 GB

\$19.12/hour

• Available now

B200 SXM x 1

180GB VRAM
24 vCPUs · 200 GB

\$5.00/hour

• Available now

Launch

보정/정제

경계보정

한국어 특성상 어미/조사/호
칭 등등 존재 가능.

최한림님께 -> 최한림

동일 PII 후보 정리

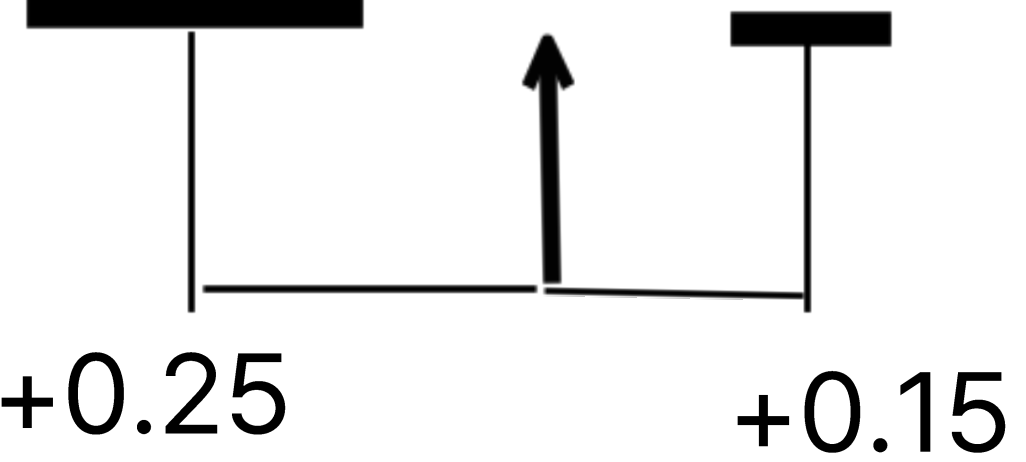
3개의 서로 다른 디텍터가 존재하기 때문에
비슷한 PII 후보를 정할 수 있음.

점수 기반 문맥 보정

앞뒤 +- 1단어

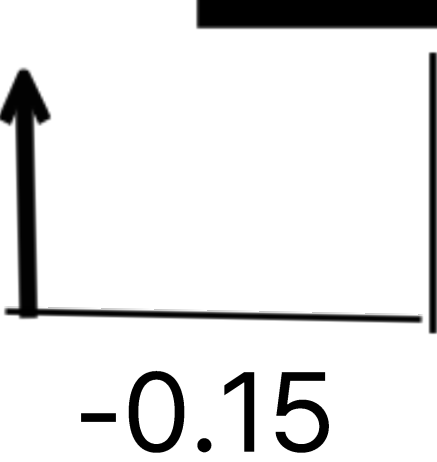
문맥보정 - 오탐과 미탐 감소

"고객명 김하늘 님, 연락처 010-1234-5678입니다"



PII

"오늘 하늘이 참 맑네요"



PII : X

Test Guardrails:

korean-pii

Test guardrail and compare results

Input Text ⓘ

📋 Copy Input

안녕하세요, 김민수입니다. 주민번호는 020814-3233139, 전화 010-2345-6789, 서울시 강남구 테헤란로 152에 살고, 이메일은 minsu.kim@gmail.com 입니다.

Press Enter to submit • Shift+Enter for new line

Characters: 106

Results

✓ korean-pii

🕒 2138ms

📋 Copy

Output Text

안녕하세요, [PERSON_1]입니다. 주민번호는 [REDACTED], 전화 [PHONE_1], [ADDRESS_1]에 살고, 이메일은 [EMAIL_1] 입니다.

Characters: 90

Success Env: default 

Request & Response

Pretty JSON

^ ... Input Tokens: 87 Cost: \$0.000218 

USER

다음 고객 상담 내용을 3줄로 요약해줘.
상담사: 안녕하세요 무엇을 도와드릴까요?
고객: [PERSON_1]입니다. 전화번호는 [PHONE_1] 이고요, [ADDRESS_1]에 삽니다. 이메일은 [EMAIL_1] 이에요. 주문한 상품이 아직 안 왔어요.

^ ✨ Output Tokens: 65 Cost: \$0.000650 

ASSISTANT

고객 [PERSON_1]님이 주문한 상품이 아직 배송되지 않았다고 문의하셨습니다. 고객의 연락처 정보는 [PHONE_1], [EMAIL_1]이며, 주소는 [ADDRESS_1]입니다. 상담사는 고객의 문제를 해결하기 위한 조치를 취할 것입니다.

🔍 평가 결과: 핵심 성능 요약

Summary of Key Indicators



종합 탐지 성능

92.4%



마스킹·차단 판단 성능

94.6%



고위험 정보 탐지율

100%



한국어 조사 분리 정확도

99.89%



개인정보 원문 로그 노출

0건

최종 배포 형태



korean_pii_guardrail_v0_2

 INTEGRATIONS.md	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 README.md	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 __init__.py	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 apisix.md	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 asgi_middlewares.py	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 azure_apim_policy.xml	feat(adapters): multi-gateway adapter layer for the PII guardrail
 bedrock.py	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 cloudflare_worker.js	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 core.py	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 envoy.md	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 guardrails_ai.py	feat(adapters): multi-gateway adapter layer for the PII guardrail
 kong.md	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 langchain.py	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 litellm_hook.py	feat(adapters): multi-gateway adapter layer for the PII guardrail
 nemo_guardrails.py	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 openai_wrapper.py	feat(adapters): more gateways/frameworks — LangChain, OpenAI/Bedrock ...
 portkey_webhook.py	feat(adapters): multi-gateway adapter layer for the PII guardrail



한학기동안 고생하셨고 감사합니다

