한국어 LLM Gateway에서 변이형 PII 우회 저감을 위한 LLM-free Layer 0 Guardrail 설계와 평가

양유상[1]*, 김민우[1], 정연서[1], 이병천[1], 임정묵[1], 최한림[1]
중부대학교[1]

Design and Evaluation of an LLM-Free Layer 0 Guardrail for Reducing Mutated PII Evasion in Korean LLM Gateways

YuSang Yang[1]*, Minwoo Kim[1], Yeonseo Jung[1], Byungchun Lee[1], Jeongmook Im[1], Hanrim Choi[1]
Joongbu University[1]

요약 : 한국어 LLM Gateway에서는 자모·Unicode·표기 변형으로 텍스트형 PII 탐지가 우회될 수 있다. 본 논문은 한국어 정규화와 PII 사전/패턴 탐지를 결합한 LLM-free Layer 0를 pre-call 계층으로 제안한다. 10k payload에서 L0-L3는 pre-LLM TPR 94.32%, KR-sem TPR 96.39%였고, L1-L4 full-stack은 90.96%, 87.40%였다. synthetic clean-pass 10k는 FP 0건, local L0 단독 p99 평균은 0.407ms였다. 결과는 변이형 PII의 저지연 사전 차단 가능성을 보인다.

Key Words : Korean PII(한국어 개인식별정보), LLM Gateway(LLM 게이트웨이), Layer 0 Guardrail(Layer 0 가드레일), Korean Normalization(한국어 정규화), Pre-Call Blocking(사전 차단)

1. 서 론

대규모 언어모델(LLM)을 활용한 서비스에서는 사용자 입력과 모델 출력에 개인정보가 포함될 수 있다. LLM Gateway는 여러 모델 호출이 집중되는 경로이므로, 개인정보를 사전에 차단하거나 마스킹하는 공통 보안 계층이 필요하다. 그러나 한국어 개인정보는 주민등록번호·전화번호 같은 정형 식별자뿐 아니라 진단명, 알레르기, 종교, 직위, 소속, 결혼상태 등 개인과 결합될 경우 식별 또는 민감정보 노출 위험을 높이는 텍스트형·의미형 속성을 포함한다. 본 논문에서는 이를 운영적 PII(operational PII)로 정의하고, 정형 개인식별자와 민감·개인속성 표현을 함께 탐지 대상으로 삼는다. 이러한 값은 자모 분해, 호환 자모, Unicode 변형, 초성 표기, 야민정음, 로마자 혼합, 구분자 삽입으로 표면형을 쉽게 바꿀 수 있어 범용 탐지기를 우회할 수 있다.

기존 PII 탐지는 정규식, NER(Named Entity Recognition, 개체명 인식), 관리형 보안 필터, LLM judge를 활용한다. 그러나 범용 탐지기는 한국어 변이형 텍스트 PII에 충분히 최적화되어 있지 않을 수 있고, LLM judge는 의미 판별 능력을 제공하지만 외부 API 비용과 지연이 추가로 발생한다. 본 논문은 한국어 LLM Gateway 앞단에서 동작하는 LLM-free Layer 0 Guardrail을 설계하고, 10,000건(10k) payload와 추가 synthetic clean-pass 10,000건 평가를 통해 한국어 PII 사전 차단 효과를 검증한다.

2. 관련 연구 및 위협모델

한글 정규화 연구는 유니코드 환경에서 완성형 음절, 조합형 자모, 호환 자모가 혼재할 때 원래 의도와 다른 조합이 생길 수 있음을 지적하였다[1]. 한국어 PII 연구는 개인식별번호와 개인정보 주석 체계를 다루었으나, LLM Gateway의 pre-call 계층에서 한국어 표기 변형, 지연, 비용, 오탐을 함께 평가한 연구는 제한적이다[2]. 문자 수준 공격 연구는 invisible character, homoglyph, reordering 등 인코딩 교란이 NLP 시스템을 공격할 수 있음을 보였고[3], 최근 guardrail robustness 연구도 typo, camouflage, cipher, veiled expression 등의 mutation이 가드레일 탐지를 약화시킬 수 있음을 보고하였다[4]. 한국어의 자모·음운 기반 변형 가능성은 별도의 한국어 위협모델을 요구한다[5].

본 연구의 공격자는 시스템 권한을 침해하지 않고 입력 문자열만 변형할 수 있다고 가정한다. 공격 목표는 실제 PII를 포함한 요청이 BLOCK 또는 MASK 없이 LLM까지 전달되도록 하는 것이다. 방어자는 정상 한국어 입력의 통과율을 유지하면서 변이형 PII를 pre-call 단계에서 탐지해야 한다. 반복적인 우회 시도는 단순 오입력이 아니라 개인정보 검사 회피 목적의 잠재적 보안 이벤트로 간주한다.

3. 연구내용

3.1 제안 방법 개요

제안 Layer 0는 사용자 입력이 외부 guardrail 및 target LLM으로 전달되기 전에 실행되는 로컬 pre-call 계층이다. Layer 0는 normalizer와 PII detector로 구성된다. Normalizer는 1.Unicode 호환 문자 및 폭(width) 변형 정규화, 2. 조합형·호환 자모의 한글 음절 복원, 3. invisible/control character 제거, 4. 반복 구분자와 공백 collapse, 5. 숫자·기호 표기의 canonicalization을 순차 적용한다. Detector는 원문 문자열과 정규화 문자열을 모두 검사하며, 주민등록번호·전화번호·계좌번호 등 정형 PII는 패턴 기반 규칙으로, 진단명·알레르기·종교·직위·소속 등 KR-sem 항목은 한국어 키워드/의미형 사전과 문맥 규칙으로 판정한다. 정책은 PASS, MASK, BLOCK으로 구분하며, 식별자 값이 보존 가능한 경우 MASK를 우선 적용하고, 민감속성 또는 우회 의도가 강한 변이형 입력은 BLOCK으로 처리한다. 이 방식은 LLM judge를 대체하기보다 judge 이전에 저비용·저지연 선행 방어선을 두는 cascade 구조이다.

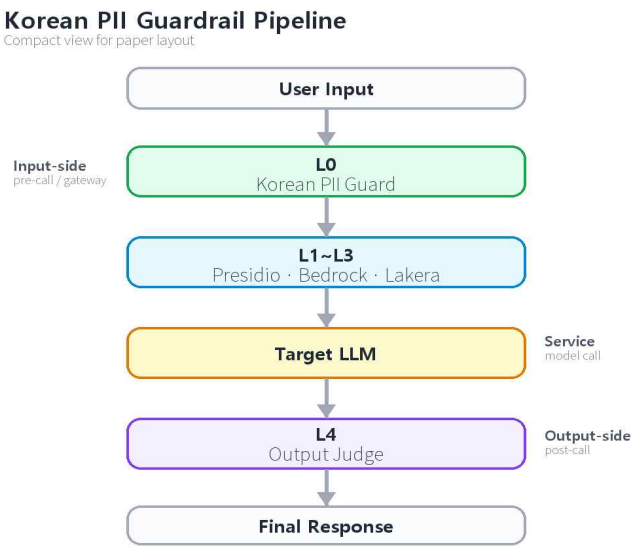

Fig. 1. Architecture of Layer 0 Pre-Call Guardrail in a Korean LLM Gateway

3.2 실험 설계

Table 1. Evaluation Configurations

Config	Stack	Role
A	L1-L3	Baseline without LLM judge
B	L1-L4	Baseline with GPT-4o-mini judge
C	L0-L3	Proposed Layer 0 without LLM judge
D	L0-L4	Full stack with Layer 0 and LLM judge

평가는 Table 1의 네 가지 구성으로 수행하였다. A는 기존 L1-L3 baseline, B는 A에 GPT-4o-mini 기반 L4 judge를 추가한 구성, C는 제안 Layer 0를 L1-L3 앞단에 배치한 구성, D는 Layer 0와 L4 judge를 모두 포함한 구성이다. 평가 데이터는 ground-truth PII를 포함한 positive payload 10,000건과 PII가 없는 synthetic clean-pass 10,000건으로 구성하였다. Positive payload는 정형 PII, KR-sem PII, 정형·의미형 혼합 PII를 포함하며, 자모 분해, 호환 자모, Unicode/폭 변형, 초성 표기, 야민정음, 로마자 혼합, 구분자 삽입 변이를 포함하도록 구성하였다. Clean-pass corpus는 개인정보가 포함되지 않은 일반 질의와 업무·보안·의료·금융 도메인 정상 질의를 포함하여 오탐 여부를 확인하는 데 사용하였다.

지표는 pre-LLM TPR, full-stack TPR, LLM exposure, KR-sem TPR, p99 latency로 구분하였다. Pre-LLM TPR은 target LLM 호출 전 L0-L3 단계에서 ground-truth PII payload가 TRUE detection으로 판정되어 BLOCK 또는 MASK 처리된 비율이다. Full-stack TPR은 post-call L4 judge까지 포함한 전체 파이프라인에서 ground-truth PII payload가 TRUE detection으로 판정되어 BLOCK 또는 MASK 처리된 비율이다. Full-stack TPR은 post-call L4 judge까지 포함한 전체 파이프라인에서 BLOCK 또는 MASK가 발생한 비율이다. LLM exposure는 target LLM 이전 단계에서 BLOCK/MASK 없이 통과한 비율로 정의한다. 따라서 B와 D처럼 L4가 포함된 구성의 full-stack TPR은 개인정보가 target LLM에 노출되지 않았음을 직접 의미하지 않는다. A-D의 full-stack recall과 latency는 저장된 동일 10k per-case 로그를 재집계했으며, local Layer 0 latency는 HTTP 또는 관리형 API 없이 normalizer와 detector만 5회×10,000건 반복 측정하였다.

3.3 실험 결과

Table 2. Pre-LLM Protection and Full-Stack Blocking Results on 10k Positive Payloads

Config	Pre-LLM TPR	Full-stack TPR	LLM Exposure	KR-sem TPR	End-to-end p99 Latency (ms)
A: L1-L3	80.15%	80.15%	19.85%	49.62%	1,317
B: L1-L4	80.15%*	90.96%	19.85%*	87.40%	4,819
C: L0-L3	94.32%	94.32%	5.68%	96.39%	830
D: L0-L4	94.32%*	97.23%	5.68%*	98.85%	4,762

*B와 D의 Pre-LLM TPR 및 LLM Exp.는 L4 이전 동일 구성(A 또는 C)의 값을 기준으로 해석한다. Full-stack TPR과 KR-sem TPR은 L4가 포함된 경우 post-call 탐지까지 포함한다.

실험 결과 C(L0-L3)는 pre-LLM 단계에서 TPR 94.32%, LLM exposure 5.68%를 보였다. 반면 A(L1-L3)는 pre-LLM TPR 80.15%, LLM exposure 19.85%였고, B(L1-L4)는 L4까지 포함한 full-stack TPR이 90.96%였으나 L4 이전 pre-LLM 기준으로는 A와 동일하게 TPR 80.15%, LLM exposure 19.85%로

해석된다. 따라서 C의 주요 효과는 L4 judge 없이도 target LLM 노출 위험을 A/B의 19.85%에서 5.68%로 낮춘 데 있다. D(L0-L4)는 full-stack TPR 97.23%, KR-sem TPR 98.85%로 가장 높은 탐지율을 보였으며, 이는 Layer 0와 L4 judge가 상호보완적으로 작동함을 의미한다.

KR-sem 항목에서 C는 TPR 96.39%를 보여 B의 87.40%보다 8.99%p 높았다. 동일 10k positive payload에서 B와 C의 TRUE detection outcome을 paired sample로 비교한 McNemar 검정 결과, B만 성공한 case는 b=291, C만 성공한 case는 c=627이었으며 p=2.04e-28로 유의하였다. C의 p99 latency는 830ms로 B의 4,819ms보다 낮았고, 이는 L4 judge 없이 L0-L3 전단에서 TRUE detection을 확정하는 early blocking 효과가 tail latency에 반영된 결과로 해석된다.

추가 검증에서 synthetic clean-pass 10k는 FP 0건이었고, rule-of-three 기준 95% 상한은 0.0300%였다. Local Layer 0 단독 latency는 5회×10,000건 반복 측정에서 p99 평균 0.407ms였다. KR-sem 세부 항목에서는 prescription(+46.16%p), allergy(+30.77%p), degree(+28.95%p), company(+26.19%p), blood(+26.00%p)에서 C가 B보다 크게 높았고, transaction, loan, family 및 일부 Unicode·구분자 변이에서는 L4 judge가 보완 효과를 보였다. Ablation에서 정규화 단독 gain은 0.31%p로 제한적이었으므로, 본 연구의 주장은 정규화 단독 성능이 아니라 정규화-우선 구조가 한국어 PII 사전/패턴 탐지를 유효하게 만든다는 점이다.

3.4 논의

D(L0-L4)가 full-stack TPR 97.23%로 가장 높은 탐지율을 보인 점은 Layer 0와 L4 judge가 경쟁 관계가 아니라 상호보완 관계임을 보여준다. 따라서 본 결과는 L4 judge의 제거가 아니라 L4 호출 대상을 줄이는 운영 전략으로 해석해야 한다. 실운영에서는 Layer 0를 기본 방어선으로 두고, L0-L3에서 BLOCK 또는 MASK가 확정된 요청은 L4 호출을 생략하며, 확정 판정이 나지 않은 고위험·불확실 케이스에만 L4 judge를 호출하는 Smart Skip 정책이 적합하다. C에서 94.32%가 L0-L3 단계에서 확정 처리되었지만, D가 C보다 full-stack TPR 2.91%p 높으므로 모든 요청에서 L4를 제거하는 방식으로 해석해서는 안 된다.

정책적으로는 PII 검사를 우회할 목적으로 문자 변이를 반복 입력하는 행위를 잠재적 공격으로 분류할 수 있다. 다만 과도한 개인정보 수집을 피하기 위해 원문 전체 저장보다 정규화 전후 해시, 탐지 유형, 차단 사유, 계정·세션 식별자, 시간 정보 등 최소 감사 로그를 보존하는 방식이 바람직하다. 이 로그는 사후 조사와 약관 기반 제재의 근거가 되며, 보존기간과 접근권한을 제한해야 한다.

제안 방식은 특정 LLM이나 외부 API에 종속되지 않아 고객상담, 의료, 금융, 교육, 공공 민원 등 한국어 자유문 입력 서비스에 적용할 수 있다. 다만 10k payload와 clean-pass corpus는 synthetic 중심이므로 실제 운영 로그의 문장 길이, 대화 맥락, 은어, 복합 개인정보 표현을 완전히 대표하지 않는다. 또한 외부 관리형 guardrail과 LLM judge의 latency는 API 상태와 모델 버전에 영향을 받으므로, 실제 배포 시에는 normalizer rule set, PII dictionary, detector policy 버전과 정책 변경 이력을 함께 관리해야 한다.

4. 결론

본 논문은 한국어 LLM Gateway에서 변이형 텍스트형·의미형 PII 우회를 줄이기 위한 LLM-free Layer 0 Guardrail을 설계하고 평가하였다. 10k stratified positive payload에서 C(L0-L3)는 pre-LLM TPR 94.32%, LLM exposure 5.68%, KR-sem TPR 96.39%를 보였으며, 이는 A/B의 L4 이전 기준 LLM exposure 19.85% 대비 target LLM 노출 위험을 낮춘 결과이다. 또한 D(L0-L4)는 full-stack TPR 97.23%, KR-sem TPR 98.85%로 가장 높은 탐지율을 보여 Layer 0와 L4 judge의 상호보완성을 확인하였다. Synthetic clean-pass 10k에서는 FP 0건이었고, local Layer 0 단독 p99 평균은 0.407ms였다. 이는 한국어 변이형 PII 중 상당수를 LLM 노출 이전에 deterministic하게 처리할 수 있음을 시사한다. 향후 실제 운영 분포에 가까운 clean corpus, 공격군별 ablation, 도메인별 PII 사전 확장, Smart Skip 정책의 비용·지연 절감 효과 검증이 필요하다.

참고문헌

[1] 안대혁, 박영배, “유니코드 환경에서의 올바른 한글 정규화를 위한 수정 방안,” 정보과학회논문지: 소프트웨어 및 응용, Vol. 34, No. 2, pp. 169-177, 2007.

[2] 강예지, 비립, 박서윤, 장연지, 이종규, 김한샘, “한국어 언어모델의 개인 식별 번호 처리 능력 연구,” 한국콘텐츠학회논문지, Vol. 24, No. 4, pp. 42-58, 2024.

[3] N. Boucher, I. Shumailov, R. Anderson, and N. Papernot, “Bad Characters: Imperceptible NLP Attacks,” Proceedings of the 43rd IEEE Symposium on Security and Privacy, 2022, pp. 1987-2004.

[4] E. Bassani and I. Sanchez, “On Guardrail Models’ Robustness to Mutations and Adversarial Attacks,” Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 16995-17006, 2025.

[5] B. Kim, M. Kim, H. Park, and B. Kim, “PHISH in MESH: Korean Adversarial Phonetic Substitution and Phonetic-Semantic Feature Integration Defense,” arXiv:2505.21380, 2025.