LLM 프롬프트 자동 비식별화를 위한 정책 기반 중간계층 설계

심다현, *이선영
순천향대학교
sinbi1731@naver.com, *sunlee@sch.ac.kr

Design and Implementation of a Policy-Based Middleware for Automatic De-identification of LLM Prompts

Da Hyeon Sim, *Sun-Young Lee
Dept. of Information Security Engineering Soonchunhyang Univ.

요 약

대규모 언어 모델(LLM) 활용이 확대되면서 사용자 프롬프트에 포함된 개인정보(PII)가 외부 전송, 로그, 학습 과정에서 노출될 위험이 커지고 있다. 본 논문은 LLM 호출 전 단계에서 프롬프트를 자동 비식별화하는 정책 기반 중간계층을 제안한다. 제안 구조는 위험도와 운영 환경에 따라 비식별화 강도를 동적으로 선택해 개인정보 보호와 응답 유틸리티를 동시에 확보한다.

Ⅰ. 서 론

최근 대규모 언어 모델(LLM)은 상담 자동화, 문서 요약, 분류, 질의응답 등 다양한 응용 분야에서 폭넓게 활용되고 있다. 그러나 LLM에 입력되는 프롬프트에는 이름, 전화번호, 주소, 계좌번호와 같은 개인정보(PII)가 포함될 수 있으며, 이 정보가 외부 API 전송이나 로그 저장, 모델 학습 과정에서 노출될 경우 심각한 개인정보 침해로 이어질 수 있다.

특히 외부 LLM API를 사용하는 환경에서는 입력 데이터에 대한 통제권이 제한적이기 때문에, 개인정보 보호를 위한 사전 대응이 필수적이다. 이에 본 논문은 LLM 앞단에 정책 기반 중간계층을 배치하여 프롬프트 단계에서 자동으로 개인정보를 탐지하고 비식별화하는 구조를 제안한다.

Ⅱ. 본론

기존 개인정보 보호 연구는 데이터 마스킹, 가명처리, 익명화, 차등 프라이버시 등의 기법을 중심으로 발전해왔다. 이러한 기법들은 주로 데이터 저장소 또는 학습 데이터 전처리 단계에 적용되며, 실시간 사용자 입력에 대한 동적 정책 적용에는 한계가 있다.

최근 LLM 보안과 관련된 연구에서는 프롬프트 필터링이나 출력 검열을 통해 민감 정보 노출을 방지하고자 하나, 입력 단계에서의 체계적인 정책 기반 비식별화 구조에 대한 연구는 상대적으로 부족하다. 본 연구는 정책 결정 로직과 중간계층 구조를 결합하여 기존 접근법의 한계를 보완한다.

2. 제안 기법 개요

2.1 중간 계층 개념

제안하는 정책 기반 중간계층은 사용자 애플리케이션과 LLM 사이에 위치하여 모든 프롬프트 요청을 중재한다. 본 구조의 핵심 목표는 LLM 호출 이전 단계에서 개인정보 노출 가능성을 최소화하는 것이다. 이를 위해 중간계층은 단순한 필터링 모듈이 아닌, 정책 결정 엔진과 개인정보 처리 파이프라인을 포함하는 독립적인 보안 계층으로 설계되었다.

중간계층은 프롬프트를 입력받은 후, 해당 요청이 LLM으로 전달 가능한지 여부와 전달 시 어느 수준의 비식별화가 필요한지를 판단한다. 이 과정은 자동화되어 있으며, 운영 환경에 따라 실시간으로 정책을 조정할 수 있다.

2.2 정책 결정 기준

비식별화 정책은 다차원 기준을 기반으로 결정된다. 첫째, 주민등록번호, 계좌번호, 카드번호, 인증번호(OTP), 비밀번호와 같은 고위험 개인정보가 탐지될 경우 가장 높은 우선순위를 부여한다. 둘째, 이름, 전화번호, 이메일, 주소 등 일반 개인정보의 포함 개수와 조합을 기반으로 위험 점수를 산출한다. 셋째, 프롬프트 전송 환경을 고려하여 외부 LLM API 사용 시 내부 LLM 환경보다 보수적인 정책을 적용한다. 넷째, 상담 응답 생성과 같이 문맥 유지가 중요한 경우와 단순 요약·분류와 같이 정보 보존이 중요하지 않은 경우를 구분하여 정책을 차등 적용한다. 마지막으로 데이터 저장 여부, 사용자 권한, 조직 내부 규정 및 컴플라이언스 요구사항을 종합적으로 반영한다.

3. 정책 기반 비식별화 설계 및 구현

3.1 정책 단계 정의

정책 결정 결과에 따라 네 가지 비식별화 단계 중 하나가 선택된다. BLOCK 단계에서는 고위험 개인정보가 포함된 요청에 대해 LLM 호출을 차단하고 사용자에게 수정 안내를 제공한다. STRICT 단계에서는 고위험 개인정보를 완전 마스킹하고, 주소나 날짜와 같은 준식별자는 일반화한다.

BALANCED 단계는 상담 업무와 같이 문맥 유지가 필요한 상황을 고려하여 이름이나 고객번호를 가명처리하고, 나머지 개인정보는 부분 마스킹한다. LIGHT 단계는 개인정보 위험이 낮은 경우로, 명확한 개인정보만 최소한으로 마스킹한다.

3.2 개인정보 탐지 모듈

개인정보 탐지 모듈은 형식 기반 탐지와 의미 기반 탐지로 구성된다. 형식 기반 탐지는 정규식을 활용하여 전화번호, 이메일, 계좌번호, 카드번호 등을 탐지한다. 의미 기반 탐지는 한국어 이름, 기관명, 지명 등을 대상으로 하며, 사전 기반 탐지와 개체명 인식(NER) 기법을 병행하여 탐지 정확도를 높인다.

탐지 결과는 개인정보 유형, 위치, 신뢰도를 포함한 구조화된 메타데이터로 관리되며, 이후 비식별화 단계에서 정책 적용의 근거로 활용된다.

3.3 비식별화 처리 방식

탐지된 개인정보는 선택된 정책 단계에 따라 마스킹, 가명처리, 일반화 방식으로 변환된다. 마스킹은 비가역적 처리로 개인정보를 완전히 제거하며, 가명처리는 동일 대화 내에서 일관성을 유지하도록 설계되어 LLM의 문맥 이해를 지원한다. 일반화는 주소나 날짜 정보를 상위 개념으로 변환하여 재식별 가능성을 낮춘다.

3.4 응답 후처리 및 감사

LLM 응답이 생성된 이후에도 개인정보가 재출력되거나 새롭게 생성될 가능성을 고려하여 응답 후처리 단계를 적용한다. 이 단계에서는 응답 내 개인정보를 재탐지하여 필요 시 추가 마스킹을 수행한다.

또한 감사 로그는 원문 데이터를 저장하지 않고, 적용된 정책 단계, 처리된 개인정보 유형 및 개수와 같은 메타데이터 중심으로 기록하여 개인정보 노출을 최소화한다.

Ⅲ. 결론

본 논문에서는 LLM 활용 과정에서 프롬프트에 포함될 수 있는 개인정보(PII) 노출 위험을 완화하기 위해, LLM 호출 이전 단계에서 자동 비식별화를 수행하는 정책 기반 중간계층 구조를 설계·구현하였다. 제안 구조는 고위험 PII 존재 여부, PII 포함량, 전송 환경(외부 API/내부 LLM), 사용 목적, 저장 여부, 사용자 권한 및 컴플라이언스 요구사항을 종합적으로 고려하여 BLOCK/STRICT/BALANCED/LIGHT 중 적절한 단계를 동적으로 선택한다.

또한 형식 기반 탐지(정규식)와 의미 기반 탐지(사전/NER)를 결합해 다양한 PII를 식별하고, 정책에 따라 마스킹·가명처리·일반화를 적용함으로써 개인정보 보호와 문맥 유틸리티의 균형을 확보한다. 아울러 응답 후처리 및 메타데이터 중심 감사 로그를 통해 재노출 가능성을 낮추고 운영 측면의 추적 가능성을 제공한다.

ACKNOWLEDGMENT

본 연구는 정부의 재원으로 한국연구재단의 지원을 받아 수행한 연구임 (NO. RS-2024-00346749)

참 고 문 헌

[1] A. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] OpenAI, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.

[3] C. Dwork, “Differential Privacy,” Automata, Languages and Programming, Springer, 2006.

[4] P. Samarati and L. Sweeney, “Generalizing Data to Provide Anonymity when Disclosing Information,” Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 1998.

[5] L. Sweeney, “k-Anonymity: A Model for Protecting Privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557 - 570, 2002.

[6] National Institute of Standards and Technology (NIST), “Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),” NIST Special Publication 800-122, 2010.

[7] European Union, “General Data Protection Regulation (GDPR),” Official Journal of the European Union, 2016.

[8] 개인정보보호위원회, 「개인정보 비식별 조치 가이드라인」, 대한민국, 2020.