기업의 안전한 생성형 AI 활용을 위한 내부 데이터 유출 방지 기술 동향 연구

명재민[1], 김원빈[2], 서대희[3]
[1]상명대학교 핀테크전공 학부생
[2]상명대학교 지능데이터융합학부 박사후연구원
[3]상명대학교 지능데이터융합학부 교수
202110793@sangmyung.kr, binnaa29@naver.com, daehseo@smu.ac.kr

A Study on Data Leakage Prevention Technologies for the Secure Use of Generative AI in Enterprises

JaeMin MYEONG[1], Won-Bin KIM[2] ,Daehee SEO[3]
[1]Dept. of Fintech, Sangmyung University
[2,3]Faculty of Artificial Intelligence and Data Engineering, Sangmyung University

요 약

AI 기술의 발달과 생성형 AI의 등장은 기업의 생산성을 높이는데 기여하며, 비즈니스 모델에 새로운 패러다임을 일으켰다. 그러나 생성형 AI는 중요 데이터와 일반 데이터를 구분할 수 없어 기업의 핵심 정보와 같은 치명적인 정보가 외부에 유출될 수 있는 문제가 발생한다. 이러한 이유로 많은 기업이 생성형 AI 사용을 제한하고 있으나, 생성형 AI로 얻을 수 있는 경제적 이점을 포기하기 어려운 상황이다. 따라서, 본 연구는 안전한 데이터 관리 및 유출 방지를 위해 보안 위협과 이를 해결할 기술을 분석한다.

1. 서론

생성형 AI(Generative Artificial Intelligence)의 등장으로, 복잡한 질문이나 요구를 통해 원하는 결과물을 얻는 것이 가능해졌다. 이러한 이유로 기업들은 업무 효율성과 생산성을 높이기 위해 생성형 AI를 활용하고 있으며, 미래 비즈니스 모델에서 중요한 역할을 할 것으로 예상된다.

그러나 생성형 AI는 민감 데이터를 스스로 구분할 수 없어 기업 업무에 생성형 AI를 활용하는 과정에서 기업의 핵심 정보가 유출되는 문제가 발생한다. 이로 인해 많은 기업에서는 업무에 생성형 AI 사용을 금지하고 있으나, 업무에 생성형 AI를 사용 시 얻을 수 있는 경제적 효과가 큰 것으로 보고되고 있다[1]. 따라서, 생성형 AI를 완전히 사용하지 못하도록 규제하는 것이 아닌 생성형 AI를 사용하면서도 내부 정보 유출을 방지하기 위한 방법이 필요하다.

이에 본 논문의 2장에서는 생성형 AI 내부 데이터 보안 위협을 분석하고, 이를 기반으로 3장에서는 생성형 AI 내부 데이터 유출 방지를 위한 보안 기술을 설명한 뒤, 4장에서 활용 방안 및 향후 연구 방향을 제시한다.

2. 생성형 AI 내부 데이터 보안 위협 분석

기업 내 생성형 AI 사용 시 내부 데이터의 보안을 간과할 경우, 나타날 수 있는 보안 취약성은 다음과 같다.

- **훈련 데이터 유출:** 기업에서 독자적인 인공지능 서비스를 제공하고자 할 때, 인공지능 학습 과정에서 수집된 훈련 데이터가 응답 생성 시 노출되는 문제가 존재한다. 이는 훈련 데이터에 존재하는 기업의 중요 정보가 포함될 수 있으며, 유출될 경우 기업에 막대한 피해가 발생할 수 있다.
- **기밀 유출:** 생성형 AI 입력 과정에서 기업 또는 기관의 기밀 정보를 제공하게 될 경우, 이러한 정보가 응답 생성 시 다른 사용자에게 유출될 수 있다. 생성형 AI는 사용자의 입력으로 학습을 진행하는데, 이때 학습된 내용을 바탕으로 새로운 사용자의 질문에 답변하기 때문이다. 따라서, 생성형 AI에 기밀 정보를 입력할 경우, 비인가된 사용자가 이를 노리고 프롬프트 입력을 통해 기밀 정보를 획득할 수 있다.
- **데이터베이스 해킹 및 회원 추론 공격:** 대규모 언어모델의 경우, 사용자의 질문과 정보가 데이터베이스에 저장되기 때문에, 데이터베이스에 접

근 가능한 사용자가 불법적으로 내부 데이터를 외부로 유출하는 문제가 발생할 수 있다. 또한, 외부 공격자의 데이터베이스 불법적 접근 또는 회원 추론 공격으로 내부 데이터의 침해가 발생될 수 있다. 따라서 이러한 내부 데이터를 대상으로 하는 공격을 통해 기업 내에서 활용하는 AI 모델로부터 기업 인사 정보나 개인정보를 획득할 수 있다.

3. 생성형 AI 내부 데이터 유출 방지를 위한 보안 기술

본 장에서는 2장에서 제시한 보안 위협으로부터 데이터의 안전성을 보장하기 위한 보안 기술을 설명한다.

- **차등 프라이버시(Differential Privacy):** 개인 데이터가 포함된 데이터 세트(Data Set)에 임의의 노이즈(Noise)를 삽입함으로써 개인 데이터가 제 3자에게 식별되지 않도록 보호하는 기법을 의미한다[2]. 차등 프라이버시는 인공지능 모델에 적용되어 민감 데이터의 유출을 막고, 악의적인 사용자가 인공지능의 출력값을 바탕으로 정보를 조합하여 기업 내부 데이터에 접근하는 것을 차단한다. 또한, 노이즈가 내부 데이터의 통계값에 영향을 주지 않도록 설계함으로써 인공지능의 성능은 유지하면서 데이터의 보안성을 높일 수 있다.
- **언러닝(Unlearning):** 학습된 인공지능 모델에서 재학습을 하지 않고 필요 없는 데이터를 훈련된 모델에서 제거하는 기술을 의미한다. 이를 통해 불필요한 데이터 노출을 방지하고, 모델이 민감한 정보를 기억하지 않도록 할 수 있다. 또한, 편집을 통해 모델의 민감 정보를 제거하기 때문에 시간과 비용이 절약되며, 데이터 삭제 요청에 대한 법적 요구사항을 신속하게 처리하고 인공지능 규제에 빠르게 적용할 수 있다[3].
- **Secure MPC(Secure Multi-Party Computation):** Secure MPC를 사용하면 여러 당사자가 개인 데이터를 공개하지 않고도 공통의 관심사인 함수를 계산할 수 있어 학습과 추론에 이용되어 개인 데이터를 보호한다[4]. 또한, 학습 데이터의 기밀성과 모델의 출력값에 대한 무결성을 보장하여, 악의적인 사용자의 데이터 삽입에 의한 모델 오염 공격(Poisoning Attack)에 대처할 수 있다.
- **기밀 AI(Confidential AI):** 기밀 AI의 핵심 요소는 모델 내부 데이터의 보호와 신뢰할 수 없는 당사자 간의 신뢰 시스템을 구축하는 것이 목표이다. 이를 위해 특수 목적으로 제작된 하드웨어와 관련 펌웨어로 신뢰 실행 환경(TEE, Trusted Execution Environment) 구축을 통해 학습 데이터의 보안을 제공한다[5]. 기밀 AI는 모델의 학습 데이터 보안을 높이고, 인공지능 보안 관련 기술과 함께 적용되어 생성형 AI의 올바른 활용이 가능하다.

4. 결론

본 연구에서는 기업 환경에서 생성형 AI를 적용하였을 때, 내부 데이터의 불법적 유출의 보안 위협과 대응 기술에 대한 분석을 수행하였다.

각 기술의 활용 방안 및 기대 효과로, 차등 프라이버시는 AI 기반 데이터 분석에서 기업 고객의 개인정보를 안전하게 보호할 것으로 기대된다. 또한, 사용자의 서비스 탈퇴에 의한 데이터 삭제 요청의 경우 언러닝을 통해 신속하게 처리함으로써, 기업 고객의 개인정보 침해 위험을 최소화할 수 있다. Secure MPC는 기업 간 협업 시 데이터 노출 없이 모델 학습과 AI 성능 향상을 가능하게 하며, 기밀 AI를 이용한 하드웨어 기반의 신뢰 실행 환경 구축을 통해 안전한 AI 활용이 가능한 업무 환경을 조성할 수 있다.

이를 바탕으로 기업 내 생성형 AI 서비스를 한층 더 안전하게 사용할 수 있는 기반을 제공할 수 있을 것으로 기대한다. 향후 본 기술들을 적용한 내부 데이터의 불법 유출 방지 서비스에 대한 연구를 진행하고자 한다.

사사표기

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음 (IITP-2024-RS-2024-00438056)

참고문헌

[1] Nielsen Norman Group. "AI Tools Boost Productivity, but Long-Term Impact on Creativity Is Uncertain." NNG, 2024.
[2] “글로벌 기업의 차등 프라이버시 기술 적용 오픈 소스 지원 현황”, 수시보고서(Privacy by Trust, Trust by Privacy), 한국인터넷진흥원, 2020.
[3] Bourtoule, Lucas, et al. "Machine unlearning," 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021.
[4] Mugunthan, Vaikkunth, et al. "Smpai: Secure multi-party computation for federated learning," NeurIPS 2019 Workshop on Robust AI in Financial Services, MIT Press, 2019.
[5] “Strengthen Data Security While Unlocking New Value with Confidential Computing Solutions,” Intel® Cloud Insider Program: Content Library, Intel Corporation, 2023.