중소기업에서의 안전한 프라이빗 LLM 도입을 위한 인프라 제안: sLLM 과 클라우드 기반으로

홍지원[1], 유은선[1], 백지윤[1], 김서영[1], 오정주[2*]
[1] 성신여자대학교 융합보안공학과 학부생
[2] 성신여자대학교 연구산학협력단 연구원

{20211107, 20200937}@sungshin.ac.kr, bjy000555@gmail.com,
{20211044, winteroot}@sungshin.ac.kr

Infrastructure Proposal for the Safe Implementation of Private LLMs in SME: sLLM and Cloud Based Approach

Jiwon Hong[1], Eunseon Ryu[1], Jiyoon Baek[1], Seoyeong Kim[1], Jungjoo Oh[2]
[1]Dept. of Convergence Security Engineering, Sungshin Women's University
[2]R&DB Foundation, Sungshin Women's University

요 약

최근 몇 년 간 대규모 언어 모델의 발전과 보급이 비즈니스 운영 통합을 가속화하고 있다. 그러나 내부 데이터 유출과 같은 문제로 많은 기업들이 보다 안전한 프라이빗 LLM 을 도입하려는 움직임을 보이고 있다. 대기업과 공공기관은 높은 비용을 부담하여 온프레미스 솔루션을 선택할 수 있으나, 중소 기업 혹은 개인에게는 예산과 기술적인 한계가 존재하기 때문에 별도의 인프라가 요구된다. 이에, 본 연구는 클라우드 서비스와 네트워크 망분리를 사용하여 중소기업이 내부 데이터를 안전하게 관리하며 LLM 을 도입할 수 있는 방안을 제시하며, RAG 모델을 통한 기술적 향상 가능성 또한 제시한다.

1. 서론

최근 몇 년 동안 대규모 언어 모델(Large Language Model, LLM)의 출현과 확산으로 인해 인공 지능 환경이 새로운 전환점을 맞이하고 있다. 이러한 인공 지능 모델의 기술적 발전, 향상된 정교함 및 성능 등으로 인해, 비즈니스 운영 구조에 대한 통합이 가속화되고 있다. 대표적으로 OpenAI 社의 ChatGPT 와 같이, 공개적으로 사용 가능한 방대한 온라인 데이터로 학습된 퍼블릭 LLM 모델은 여러 기관이나 개인이 손쉽게 접근할 수 있는 서비스도 증가하고 있는 추세이다. 그러나, 방대한 퍼블릭 LLM 에서 기업 내부 데이터 유출 등이 발생한 사례들로 인해 최근 기업들은 프라이빗 LLM 도입이 확대되고 있다[1].

대기업 및 공공기관에서는 기업 데이터를 관리하기 위해 온프레미스(On-Premise) 방식을 택하고 있으며, 기관의 자체적인 기술력을 기반으로 한 프라이빗 LLM 을 도입함으로써 보안을 강화한다. 그러나 온프레미스 방식은 설치 및 유지보수 비용이 높은 가격으로 책정되어, 중소기업에서는 도입하기 어려운 현실이다[2].

따라서, 본 연구에서는 차세대 핵심 기술인 LLM 을 중소기업에서도 안전하게 도입하기 위한 방안을 제시한다. 클라우드 서비스와 망분리를 통해 물리적인 서버 설치 없이 내부 데이터와 서비스를 효과적으로 관리할 수 있으며, 검색 증강 모델(Retrieval-Augmented Generation, RAG)을 이용하여 sLLM 의 기술적 향상 또한 고려하였다.

2. 배경

최근 대부분의 기업 및 기관은 인터넷 기반에서 내부 업무 활동을 진행하고 있다. 기업의 영리를 위한 영업 활동과 고객 서비스 등의 대부분이 인터넷과 모바일 환경에서 이루어짐에 따라, 기업과 기관을 대상으로 한 사이버 공격이 증가하고 있다.

따라서 현재 모든 공공기관, 일정규모 이상의 금융기관, 방산업체 등은 기관/기업 내 네트워크 환경을 인터넷망과 업무망을 분리하여 사용하도록 의무화하고 있다. 특히 공공기관은 매년 정보보호관리 실태평가 감사를 받으며 망분리가 안 되었을 경우, 낮은 점수를 부여받고 있다[3].

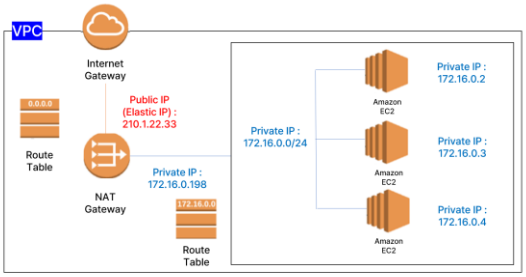
또한 최근 대규모 언어 모델이 야기할 수 있는 환

각 현상을 최소화하는 검색 증강 모델인 RAG 모델이 급부상하고 있다[4]. RAG 모델은 감사 가능한 최신 정보를 제공함으로써 환각 문제를 해결할 수 있는 방안을 다음과 같이 제공한다. 학습 데이터 외 외부 데이터에 대한 접근이 가능하여, 제공되는 정보의 신뢰도가 향상된다. 뿐만 아니라 최신 정보만을 다루기 때문에, 기업 내부에서 사용하는 sLLM(Small LLM)의 작은 파라미터 값으로 인해 부정확한 답변이 출력되는 것을 최소화할 수 있어 많은 기업들이 사용하고 있다.

따라서 본 논문에서는 기업의 기밀 데이터가 유출되는 것을 최소화하기 위한 망분리와, 기업에서 사용되는 LLM 의 한계를 보완할 수 있는 RAG 모델이 결합된 인프라를 3 절과 같이 제안한다.

3. 프라이빗 LLM 인프라 도입 제안

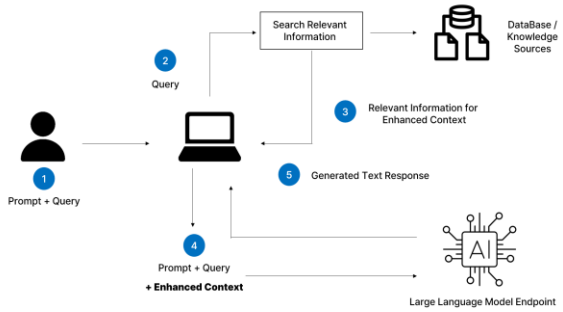
2 절에서 언급한 바와 같이, 망분리는 정보보호를 고려한 보안성 강화를 위해 필요한 기술적 조치이다.



(그림 1) AWS 기반의 망분리 인프라 제안

(그림 1)은 AWS(Amazon Web Services)를 활용하여 기업 내부망 인프라 구축을 도식화한 그림이다. EC2(Elastic Compute Cloud)는 Private Subnet 으로 IP 가 각각 할당받는 것이 가능하므로, 기업 부서 내외 등에서 프라이빗 IP 안에서의 업무 수행 시, 데이터 유출을 최소화하여 보안을 강화할 수 있다.

이를 통해 중소기업에서는 온프레미스 서버보다 경제적으로 강화된 인프라를 AWS 기반으로 구축할 수 있다.



(그림 2) RAG - LLM 결합을 통해 강화된 프레임워크

(그림 2)처럼 RAG 모델은 프롬프트와 쿼리 기반으로 LLM 이 답변을 출력하는 형태로 융합되어 사용된다. RAG 모델의 핵심은 모델 외부, 주로 벡터 데이터베이스, 지식 그래프 또는 구조화된 데이터 테이블에 주제별 전문 지식을 담는 데 있다. 이 설정은 데이터 저장소와 최종 사용자 사이에 정교하고 지연 시간이 짧은 중간 계층을 생성한다.

따라서 본 논문은 기업 특성상 sLLM 을 활용해야 하는 중소 기업에서의 안전한 프라이빗 LLM 도입을 위해 (그림 1) 기반의 인프라 내에서 (그림 2)와 같은 기술적 도입을 제안한다. 인터넷 게이트웨이부터 시작하여 방화벽으로 보호되는 구조를 통해, 인프라 구조부터 제어된 내부 사용자 접근으로 이어지는 내부망 인프라와 LLM-RAG 모델로 보안 및 성능 강화를 기대할 수 있다.

4. 결론

본 논문에서는 중소기업을 위해 적은 비용으로 안전하게 프라이빗 LLM 을 구축할 수 있는 인프라 모델을 제안한다. 보안이 강화된 프라이빗 클라우드 내부 망 분리를 통해 접근 권한에 따라 내부 데이터에 접근할 수 있도록 구성하였다. 또한, RAG 모델을 통해 환각 현상을 줄이고, 조직 내 데이터 베이스와 지식 기반으로 확장하여 정확성을 확인하며 신뢰성 있는 정보를 보장할 수 있도록 하였다.

향후 연구에서는 RAG 모델의 정교화를 기반으로 하여 모델 자체의 기능적 향상을 목적으로 해당 프레임워크를 발전시키고자 한다. 기업의 데이터가 실시간으로 업데이트되어 신속한 검색 결과를 제공하는 RAG 모델의 성능 향상은 조직 내부의 데이터베이스 구조와의 통합으로, 보다 더 정확하고 신뢰할 수 있는 정보 제공이 가능하다.

5. 참고 문헌

[1] 홍국기, "LG CNS 코드 생성형 AI 에 최적화한 LLM 개발", 연합뉴스, 2024.01.31., [Online]. Available: https://n.news.naver.com/mnews/article/001/0014478102?sid=105

[2] Malallah, Hayfaa Subhi et al, “Performance analysis of enterprise cloud computing: a review”, Journal of Applied Science and Technology Trends, Vol. 4, No. 1, pp.1-12, 2023.

[3] 백현철, 손태근, 김민아, 전정남, “망분리와 망연계 시스템의 기술 동향”, 항공우주산업기술동향, Vol. 20, No. 2, pp.120-133, 2022.

[4] Lewis, Patrick et al, “Retrieval-augmented generation for knowledge-intensive nlp tasks”, 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, 2020, pp. 9459-9474.