Headline

사이버보안 특화 Vertical AI 구축 방안

혁신사업본부 사이버보안 AI 랩스 김기남 책임

■ 1. 대규모 언어 모델(LLM) 발전과 Vertical AI 의 부상

1.1 LLM 발전 동향

대규모 언어 모델(Large Language Model, LLM)은 방대한 텍스트 데이터를 학습하여 인간의 언어를 이해하고 생성하는 능력을 갖춘 인공지능(AI) 모델이다. LLM 은 트랜스포머(Transformer)라는 혁신적인 신경망 아키텍처를 기반으로 문장의 문맥과 단어 간의 복잡한 관계를 파악한다. 이는 단순 텍스트 생성을 넘어 번역, 요약, 질의응답, 코드 생성 등 다양한 자연어 처리(NLP) 작업을 높은 정확도로 수행하도록 한다.

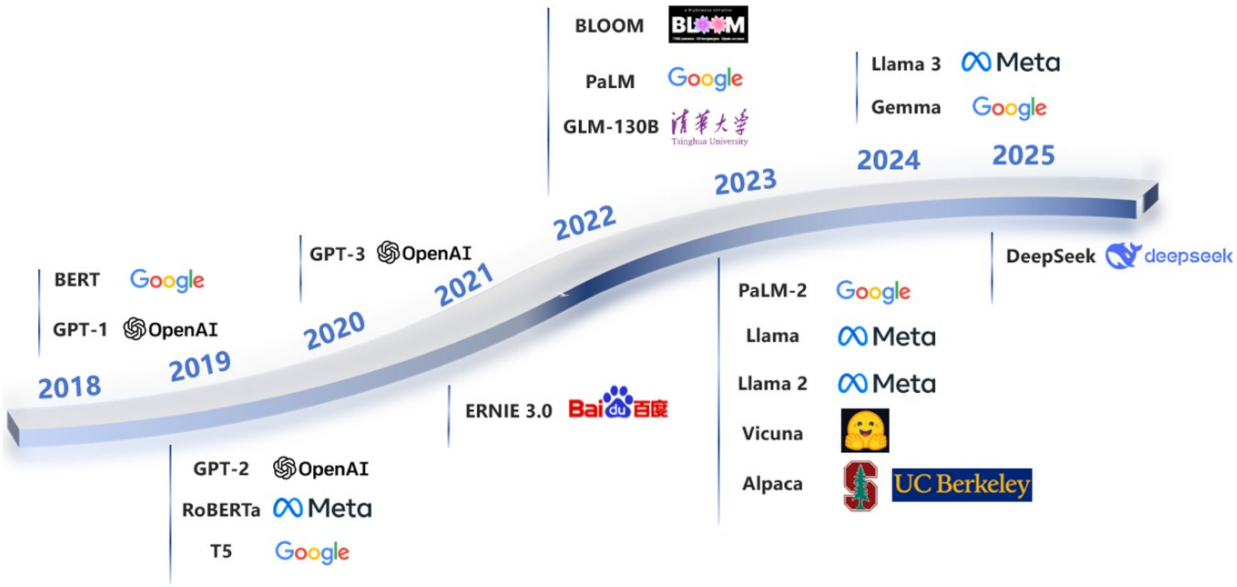

그림 1. 연도별 LLM 주요 모델

지금까지 수많은 LLM 이 개발되었다. 각 LLM 은 '스케일링 법칙(Scaling Law)'에 따라 데이터의 크기가 커질수록 성능이 비약적으로 향상된다는 점이 입증되었다. 이는 초기 LLM 이 모델 크기와 학습 데이터 양을 늘리는 '규모의 경쟁'에 집중했던 이유가 결과로 이어졌다.

지금까지 수많은 LLM 이 개발되었다. 초기에는 주로 모델 파라미터 수와 학습 데이터 규모를 확장하는데 초점을 두었다. 이는 모델의 규모와 학습 데이터 양이 커질수록 성능이 비약적으로 향상된다는 '스케일링 법칙(Scaling Law)'에 근거한 것으로, 이러한 방향성은 오랫동안 LLM 발전의 핵심 기조로 자리해왔다.

그러나 최근에는 단순히 크기를 키우는 경쟁을 넘어, 성능과 효율성을 동시에 강화하는 방향으로 빠르게 발전하고 있다. 이러한 변화의 핵심 동향은 모델 아키텍처 혁신, 컨텍스트 윈도우 확장, 그리고 추론 능력 고도화 세 가지로 요약할 수 있다.

- 모델 아키텍처 혁신

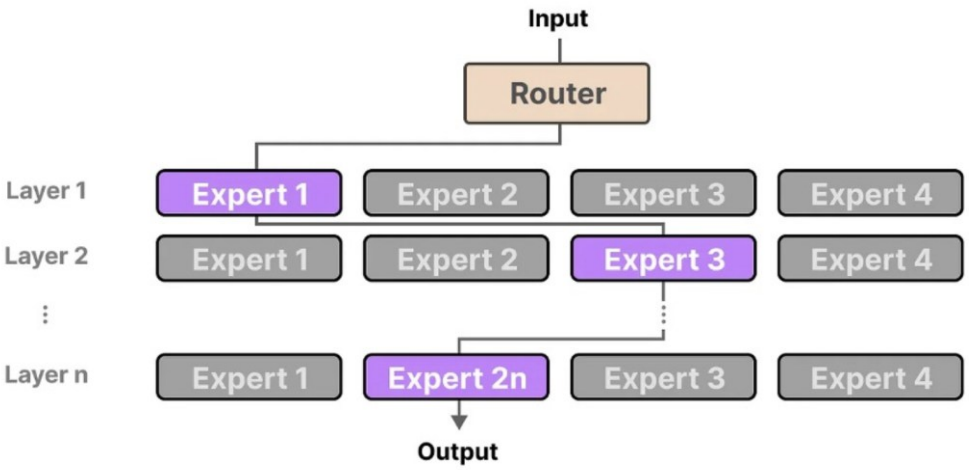

그림 2. 전문가 혼합(Mixture-of-Experts, MoE) 구조

전문가 혼합(Mixture-of-Experts, MoE) 구조는 특정 과업에 최적화된 일부 전문가 네트워크만 선택적으로 활용하여 연산 효율을 극대화한다. 최신 모델들은 MoE 구조를 채택하여, 하나의 대형 모델로 다양한 서비스와 도메인에 효율적으로 대응하고 클라우드 비용과 GPU 자원 소모를 최소화하고 있다.

- 컨텍스트 윈도우 확장

초기 LLM 은 수천 토큰 수준의 텍스트만 처리할 수 있었다. 하지만, 최근 Gemini 1.5 Pro, Llama 4 Scout 등의 모델은 수십만에서 백만 토큰 이상을 한 번에 처리한다. 이를 통해 수천 페이지에 달하는 연구 논문, 기술 문서, 법률 자료 등 방대한 문서를 단번에 분석하고 전체 맥락 속에서 핵심 정보를 정확히 찾아낼 수 있어, 전문가의 지식 활용과 비즈니스 환경의 데이터 처리 효율을 크게 향상시킬 수 있다.

- 추론 능력 고도화

최신 모델은 단계별 사고(Chain-of-Thought, CoT) 기법과 강화학습을 통해 문제 해결 과정을 논리적으로 설명하는 능력을 내재화하고 있다. GPT-4o 나 DeepSeek-R1 과 같은 모델들은 수학, 코딩, 논리 문제에서 높은 정확도를 보이며 결과와 이유를 함께 제시하여 신뢰도를 높였다. 이는 장기적으로 다단계 의사결정이 가능한 자율 AI 에이전트로의 발전 가능성을 시사한다.

1.2 LLM 의 한계와 Vertical AI 의 발전

LLM 의 급격하게 발전했음에도 불구하고 금융, 제조, 보안 등 전문 산업 분야에 AI 를 도입하는 데에는 여전히 어려움이 있다. LLM 은 방대한 지식을 바탕으로 다양한 주제에 대해 폭넓게 답할 수 있지만, 특정 산업이나 도메인에 그대로 적용하기에는 다음과 같은 한계점이 있다.

- 정확성과 신뢰성 부족

LLM 은 사실과 다른 정보를 그럴듯하게 생성하는 '환각(Hallucination)' 현상에서 자유롭지 않다. 또한 학습하지 않은 최신 데이터가 있을 시 정확성이 떨어진다. 예를 들어, 최신 제로 데이 취약점 관련 질문에 부정확한 답변을 하거나, 존재하지 않는 IP 를 공격자로 지목하는 등 잘못된 공격 패턴을 제시할 수 있는 위험성이 있다.

- 전문 지식 및 맥락 이해의 한계

분야마다 사용하는 고유 용어 등 LLM 이 이해하지 못하는 깊이 있는 전문 지식이 존재하기 마련이다. 사이버보안의 경우, 대응 시 공격 기법(TTPs), 복잡한 로그 데이터 등 깊이 있는 전문 지식이 필요하다. LLM 은 용어의 미묘한 차이나 특정 공격 벡터의 맥락을 완벽하게 이해하지 못해 위협에 효과적인 대응이 불가하다.

- 데이터 유출 및 보안 위험

LLM 은 주로 외부 API 통신을 통해 사용된다. 내부 시스템 로그, 악성코드 샘플 등 민감한 데이터를 전송하는 것은 그 자체로 심각한 데이터 유출 사고로 이어질 수 있다.

최근 이러한 한계로 특정 산업과 도메인에 최적화된 Vertical AI 의 필요성이 커지고 있다. Vertical AI 는 특정 내부 데이터와 전문 지식을 직접 학습하거나 실시간으로 참조하여 보다 정확하고 신뢰성 있는 답변을 지원한다. 또한, 통제된 내부 인프라에서 모델을 운영함으로써 데이터의 외부 유출을 차단할 수 있다.

Vertical AI 를 구현하기 위해 각 도메인에 특화된 LLM 을 체계적으로 설계하고 구축하는 것이 핵심 경쟁력으로 꼽힌다. 도메인 특화 LLM 은 단순한 기술적 구현을 넘어 산업별 문제 해결과 혁신을 가속화하는 Vertical AI 의 중심 엔진으로 자리잡고 있다.

■ 2. 사이버보안 특화 LLM 구축 방안

사이버보안 특화 LLM 은 내부 데이터와 보안 전문 지식을 기반으로 정확한 답변을 생성할 수 있는 LLM 과 그 외 시스템 구축을 목표로 한다. 구축을 위한 세 단계는 1. 기반 모델 선정 2. RAG 구축 3. LLM 파인튜닝 순이며, 이는 신입 수준의 모델을 전문가 수준으로 성장시키는 과정을 거쳐야 한다.

그림 3. 사이버보안 특화 LLM 구축 프로세스

2.1 목적에 맞는 기반 모델 선정

사이버보안 도메인 특화 LLM 을 구축하기 위해서는 우선 목적에 부합하는 기반 모델(Base Model)을 선정하는 과정이 필요하다. 단순히 최신 성능의 LLM 을 선택하는 것이 아니라, 실제 보안 환경에서 요구되는 정확성, 확장성, 보안 내재화 요소 등을 종합적으로 고려해야 한다.

기반 모델 선정 시 고려해야 할 주요 기준은 다음과 같다.

- 모델 크기와 성능의 균형

크기가 큰 모델일수록 정확도가 높지만 학습과 추론에는 많은 자원이 필요하다. 또한, 실시간성이 중요한 사이버 위협 대응 환경에서는 응답 지연(latency)은 문제가 발생할 수 있다. 따라서 업무 목적과 리소스 제약에 따라 모델 규모를 조절하는 것이 중요하다.

- 보안 특화 데이터와의 적합성

사이버보안 데이터는 일반 텍스트와 달리 로그, 코드, 스크립트, 취약점 데이터(CVE) 등 다양한 비정형 텍스트로 구성된다. 이러한 보안 특화 데이터를 효율적으로 이해하고 처리할 수 있는 모델을 선택해야 한다. SecBench, SecEval 등 사이버보안 벤치마크 데이터셋을 통해 모델을 사전에 검증한다.

- 보안 내제화

개인정보 비식별화, 민감 데이터 필터링 등 보안 기능을 제공하거나 연계가 용이한 모델인지 고려한다. 또한, 모델이 입력 데이터나 학습 데이터에 포함된 민감 정보를 외부로 유출하지 않고 안전하게 처리할 수 있는지 확인해야 한다.

기반 모델은 최대 성능만 고려하는 것이 아니라, 한정된 자원 내에서 성능, 비용, 보안 등 다양한 지표 간의 트레이드오프 관계를 이해한 후 전략적으로 선정해야 한다. 최신 정보 반영, 부족한 도메인 전문성과 같은 문제는 이후 설명할 RAG 와 파인튜닝을 통해 보완할 필요가 있다.

2.2 최신성과 신뢰성 강화를 위한 RAG 구축

LLM 은 방대한 지식을 가지고 있지만, 학습 시점이 고정되어 있어 최신 정보를 반영하지 못하거나 사실이 아닌 정보를 그럴듯하게 만들어내는 환각 현상을 보인다. 또한, 명확한 답변의 근거를 제시하지 못하면 신뢰도에 문제가 생기기도 한다.

이러한 문제를 해결하기 위해 등장한 기술이 검색 증강 검색(Retrieval-Augmented Generation, RAG)이다. RAG 는 LLM 이 더 정확하고 신뢰할 수 있는 답변을 생성하도록 돕는 기술로, LLM 이 치르는 오픈북 시험에 비유할 수 있다. LLM 이 자체적으로 학습한 지식에 의존하는 것이 아니라, 질문을 받으면 관련된 정보를 검색하여 참고한 뒤 그 내용을 바탕으로 답변을 생성하는 방식이다.

RAG 를 통해 LLM 은 답변의 근거가 되는 최신 위협 정보나 내부 데이터 소스를 제시할 수 있어, AI 판단의 신뢰성과 설명 근간을 확보할 수 있다. 또한, 기존에 파편화되어 있던 각종 위협 정보, 시스템 로그 등을 하나의 지식 베이스로 연결하여 자연어 질문 하나로 모든 정보를 탐색할 수 있게 한다.

RAG 구축 및 활용 프로세스는 다음과 같다.

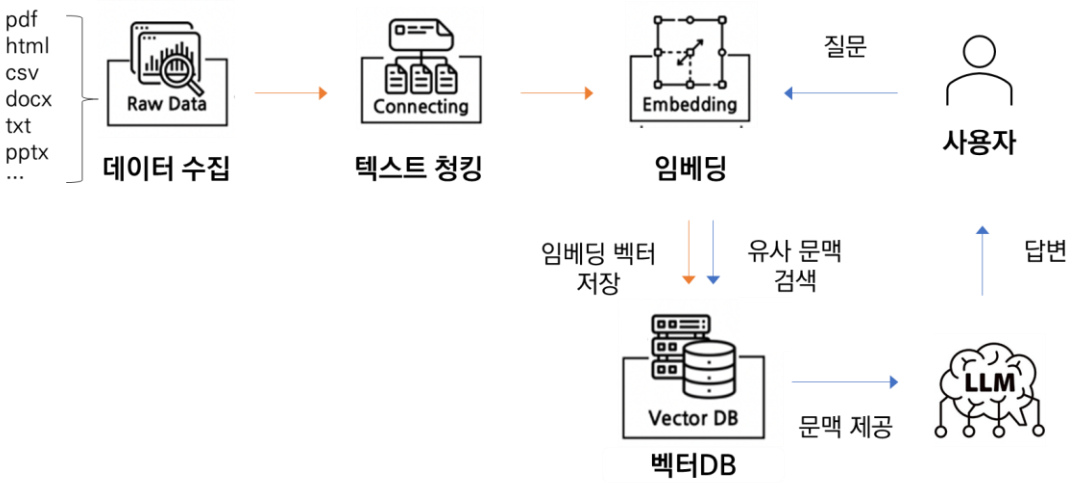

그림 4. RAG 구축 및 활용 프로세스

① 데이터 수집

답변의 근거로 삼을 수 있는 내 · 외부 데이터(pdf, docx, pptx 등)를 우선적으로 수집해야 한다. 데이터의 품질은 RAG 시스템의 성능에 직접적인 영향을 미치므로, 정확하고 최신 정보를 포함하도록 관리해야 한다.

② 텍스트 청킹

LLM 이 처리하기에 적합한 크기의 텍스트 청크(chunk) 단위로 데이터를 분할한다. 청크 크기는 LLM 의 성능, 데이터의 특성, 검색 효율성 등을 고려하여 결정한다. 너무 작은 청크는 문맥 정보를 잃을 수 있으며, 너무 큰 청크는 LLM 처리에 부담을 가중시킬 수 있다.

③ 임베딩

임베딩 모델(Embedding Model)을 사용해 텍스트 청크를 숫자로 이루어진 벡터로 변환한다. 임베딩 모델은 텍스트의 의미를 보존하면서 벡터 공간에 표현하는 역할을 한다. 고품질의 임베딩 모델을 선택하는 것이 RAG 시스템의 성능 향상에 중요하다.

④ 벡터 DB 구축

변환된 벡터는 벡터 DB(Vector Database)에 저장한다. 벡터 DB 는 유사한 벡터를 효율적으로 검색할 수 있도록 설계된 특수한 데이터베이스로 Milvus, Chroma, Weaviate 등의 벡터 DB 를 활용할 수 있다.

⑤ 검색 및 생성

사용자 질문과 유사한 문맥을 가진 데이터를 벡터 검색을 통해 추출한다. 검색된 데이터를 기반으로 LLM 이 응답을 생성한다. 이 과정에서 LLM 이 원본 데이터의 문맥을 반영하므로 정확도와 신뢰도가 높아진다.

2.3 도메인 지식 내재화를 위한 LLM 파인튜닝

RAG 를 통해 최신 데이터를 반영할 수 있으나, 특정 보안 영역에서 요구되는 심층적인 지식과 맥락 이해는 부족할 수 있다. 이를 보완하는 방법이 파인튜닝(Fine-Tuning)이다. 다만, 파인튜닝은 필수적인 것은 아니며 업무 목적과 환경에 따라 선택적으로 적용해야 한다.

파인튜닝을 통해 조직의 고유한 보고서 스타일을 따르거나 특정 공격 유형에 민감하게 반응하는 AI 를 만들 수 있다. 사이버보안 분야의 전문성을 내재화하여 보안 분석 보고서 작성, 위협 탐지 패턴 해석, 사고 대응 절차 제안 등의 업무에서 보다 정교한 답변을 제공받을 수 있다.

파인튜닝 과정은 다음과 같이 세 단계로 나타낼 수 있다.

① 데이터셋 구축

모델 학습용 데이터셋을 구축한다. 도메인 전문 데이터를 기반으로 질의응답(Question-Answer) 형식의 데이터 쌍을 여러 개 생성한다. 데이터셋의 품질은 파인튜닝 결과에 직접적인 영향을 미치므로 고품질의 데이터셋을 구축하는 것이 중요하다.

② 모델 학습

준비된 데이터셋을 사용하여 사전 학습된 기반 모델을 추가로 학습하는 단계이다. 구축된 질의응답 데이터셋을 따라 하도록 모델을 학습시키고 파인튜닝(Supervised Fine-Tuning)을 수행한다. 이 때, 모델 전체를 재학습 시키는 방식은 상당한 컴퓨팅 자원을 소모하므로 일부 파라미터만 수정하는 PEFT(Parameter-Efficient Fine-Tuning)와 같은 효율적인 기법을 사용한다.

③ 평가 및 배포

파인튜닝이 완료된 모델이 태스크에 대한 성능 향상을 달성했는지 평가한다. 또한 모델이 새로운 전문 지식을 학습하면서 기존에 가지고 있던 방대한 일반 지식을 유지하는지 검증해야 한다. 검증을 통과한 모델은 실제 운영 환경에 배포해 활용한다.

기반 모델 선정, RAG 구축, 그리고 LLM 파인튜닝의 3 단계 접근법을 통해 내부 데이터와 최신 위협 정보를 효과적으로 활용하여 정확하고 신뢰성 높은 답변을 생성하는 LLM 시스템 구축 방법을 알아보았다. 그러나 성공적인 LLM 시스템 구축만큼 중요한 것은 시스템을 안전하게 운영하고 잠재적인 위협으로부터 보호하는 것이다. 3 장에서는 LLM 시스템의 보안 취약점과 그에 대한 효과적인 대응 방안을 구체적으로 살펴본다.

■ 3. LLM 시스템 보안 취약점 및 대응 방안

LLM 이라는 최신 기술의 성장으로 기존 웹 애플리케이션과는 새로운 보안 위협들이 등장하고 있다. 금융, 의료, 보안 등 특정 산업들은 개인정보보호 및 데이터 보안에 엄격한 규제를 받고 있다. 민감한 정보가 많아 보안에 특히 주의해야 한다. 구축한 LLM 시스템에서 발생할 수 있는 보안 취약점을 사전에 파악하고 그에 대한 대응 방안을 마련하는 것이 중요해지고 있다.

3.1 LLM 시스템 보안 취약점

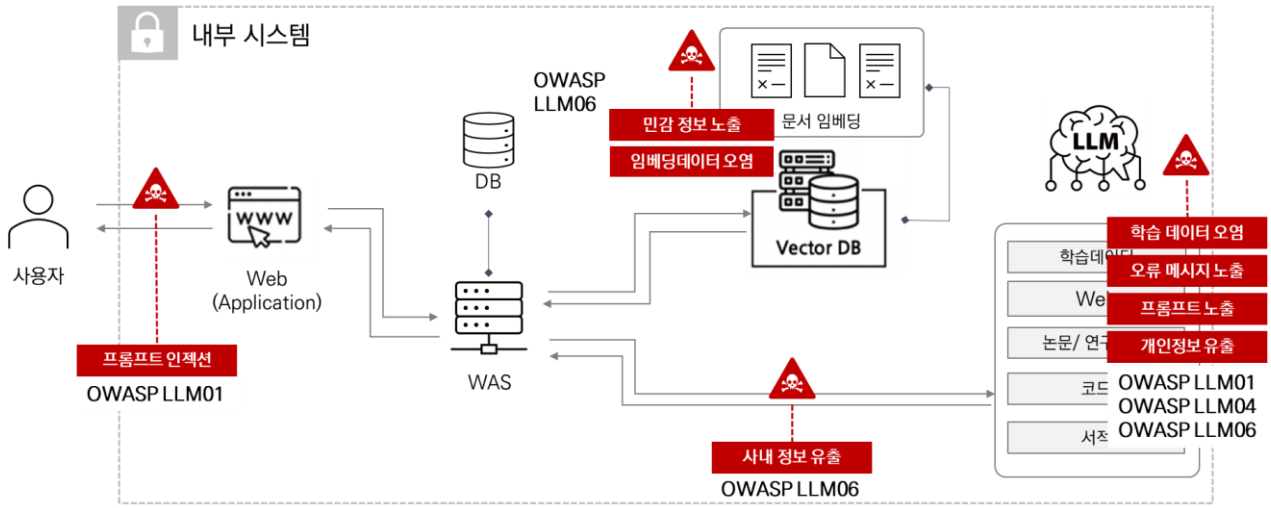

그림 5. LLM 보안 취약점

LLM 시스템에서 발생할 수 있는 취약점은 LLM 모델의 내재적 취약점과 시스템 인프라 및 운영 환경의 외재적 취약점으로 구분할 수 있다. 이 두 가지는 서로 다른 공격 경로와 방어 전략이 필요하므로 이를 분리하여 이해하는 것이 중요하다.

① LLM 자체 취약점

LLM 자체 취약점은 모델의 작동 방식과 데이터 처리 과정에서 발생하는 문제이다. 주로 모델의 예측을 조작하거나 의도하지 않은 동작을 유도하는 방식으로 나타난다. 대표적인 취약점은 프롬프트 인젝션, 민감 정보 유출, 학습 데이터 오염 등이 있다.

- 프롬프트 인젝션(Prompt Injection)

공격자가 모델에 입력하는 프롬프트를 조작하여 모델의 예측을 조작하거나 의도하지 않은 동작을 유도하는 공격이다.

- 민감 정보 유출(Sensitive Information Disclosure)

모델이 출력하는 정보 중 민감한 정보가 포함되어 있는 경우, 이를 제대로 처리하지 않아 정보가 유출되는 취약점이다.

- 학습 데이터 오염(Data and Model Poisoning)

공격자가 모델의 사전학습 또는 미세조정 과정에 사용되는 데이터셋에 악의적인 데이터를 주입하는 공격을 의미한다.

② 시스템 인프라 및 운영 취약점

LLM 을 구동하고 서비스로 제공하는 전체 시스템 환경에서 발생하는 보안 문제이다. 모델 자체보다는 모델을 둘러싼 인프라가 공격 대상이 된다. 대표적인 취약점은 공급망 공격, 인증 및 권한 관리 취약점, 임베딩 데이터 오염 등이 있다.

- 공급망 공격(Supply Chain Vulnerabilities)

LLM, 파인튜닝에 사용되는 데이터셋, 외부 라이브러리 등 LLM 시스템을 구성하는 외부 요소에서 발생하는 취약점을 통해 시스템 전체가 위험에 노출되는 취약점이다.

- 인증 및 권한 관리 취약점(Authentication & Authorization Vulnerabilities)

LLM 시스템을 사용하는 사용자의 부적절한 인증 및 권한 관리를 이용해 공격자가 시스템에 접근하거나 데이터를 조작하는 취약점이다.

- 임베딩 데이터 오염(Embedding Data Poisoning)

RAG 시스템에서 발생하는 취약점으로, 공격자가 벡터 DB 에 저장된 임베딩 데이터를 조작하는 것을 의미한다.

3.2 최신성과 신뢰성 강화를 위한 RAG 구축

LLM 시스템에서 발생할 수 있는 다양한 보안 취약점들은 시스템 설계 단계부터 체계적인 보안 아키텍처를 적용해야 효과적으로 대응할 수 있다. 사용자 인증부터 데이터 처리, 모델 응답에 이르는 전 과정에 걸쳐 보안을 내재화해야 한다.

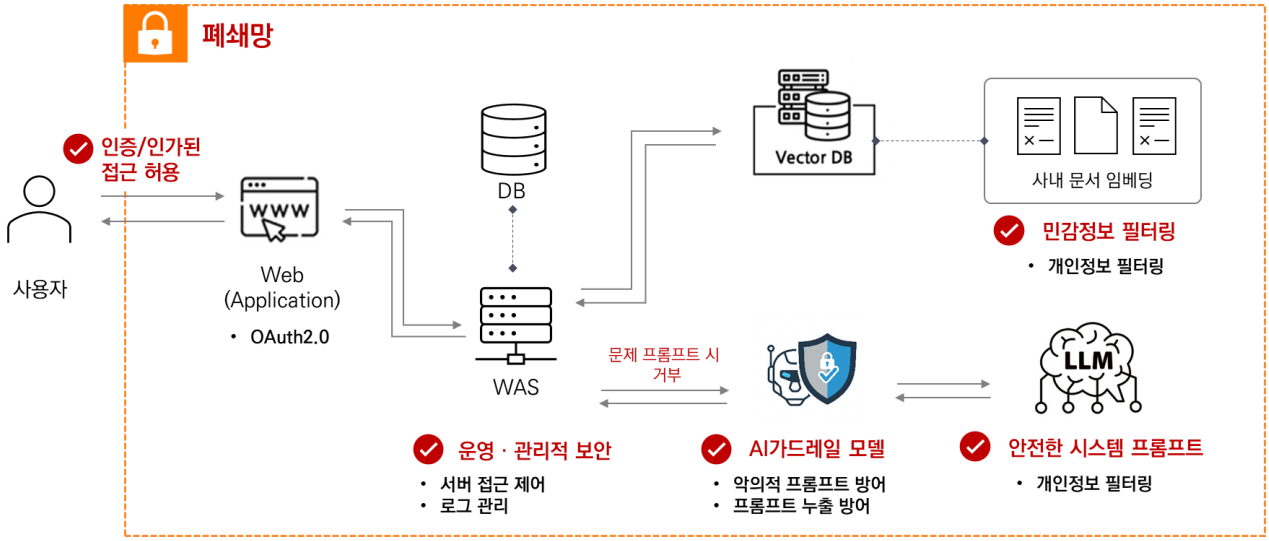

그림 6. 안전한 LLM 시스템 아키텍처

[그림 6]은 보안 취약점에 대응하기 위해 AI Guardrail 을 적용하여 입출력을 제어하였으며, 폐쇄망과 다층적 보안 구조를 통해 구축된 안전한 LLM 시스템 아키텍처 예시를 보여준다.

① AI Guardrail 을 적용한 입출력 제어

프롬프트 인젝션(Prompt Injection), 민감 정보 유출와 같이 LLM 입출력과 관련된 취약점을 대응하기 위해 입출력 내용을 필터링하는 AI 가드레일을 사용할 수 있다. 단순히 특정 키워드를 차단하는 것을 넘어, 문맥을 이해하고 정책 기반으로 잠재적 위협을 방어하는 역할을 수행한다.

- 입력 데이터 필터링

사용자의 질의는 메인 LLM 에 전달되기 전 AI 가드레일을 통해 먼저 검증된다. 이 과정에서 악의적인 공격 패턴이나 민감 정보 유출 시도가 감지되면 해당 요청을 즉시 차단하고 관리자에게 알림을 전송한다.

- 출력 데이터 제어

LLM 이 생성한 응답에 외부 유출이 금지된 민감 정보나 기밀 데이터가 포함된 경우, 이를 최종 사용자에게 전달하기 전에 차단하여 정보 유출을 방지한다.

② 폐쇄망 구성과 다층적 보안 구조

LLM 시스템은 외부와 완벽히 차단된 폐쇄망(내부망) 환경에서만 운영되도록 설계하여, 외부로의 정보 유출 경로를 원천적으로 차단한다. 또한, 모델 학습에 사용된 데이터와 실제 서비스 운영에서 처리되는 데이터를 명확히 분리하여 관리해야 한다.

- 신뢰할 수 있는 모델 공급망 확보

공급망 공격을 예방하기 위해 공신력 있는 기업이나 연구 기관이 제공하는 LLM 모델만을 채택해야 한다. 모델 배포 시에는 해시 값 검증을 통해 무결성을 확보하고, 모든 업데이트는 중앙 관리 서버를 통해서만 안전하게 수행한다.

- 강화된 사용자 접근 통제

SSO(Single Sign-On) 기반의 표준화된 인증 절차를 적용하고 역할 기반 접근 제어(RBAC, Role-Based Access Control)를 통해 사용자의 권한을 세분화하여 시스템 접근을 엄격하게 통제한다.

■ 맺음말

성공적인 사이버보안 LLM 도입을 위해서는 범용 LLM 한계를 보완한 보안에 특화된 LLM 구축에 있다. 이 과정에서 RAG, 파인튜닝과 같은 기술 구현은 물론, 프롬프트 인젝션 등의 새로운 위협에 대응할 LLM 보안 체계도 반드시 병행돼야 한다. 기술과 보안의 균형을 갖춘 특화 LLM 은 지능화되는 위협 환경에 선제적으로 대응하고, 미래 보안 경쟁력을 확보하는 핵심 동력이 될 것이다.

SK 쉴더스 사이버보안 AI 랩스에서는 AI 발전 트랜드에 따라 대규모 언어 모델(LLM)을 기반으로 한 생성형 AI 연구 과제를 진행하고 있다. 최근 자체 기술력으로 사내 내부망에서 안전하게 활용할 수 있는 '생성형 AI Shieldi(쉴디)' 서비스를 개발하여 고도화하였다. 이를 통해 사내 정책/가이드라인에 대한 맞춤형 상담과 보고서 생성 · 번역 · 요약 등 AI 기반의 업무 생산성을 높이고 Shadow AI 와 같은 잠재적 보안 위협을 최소화할 것으로 기대하고 있다.

■ 참고문헌

[1] DeepSeek-AI, DeepSeek-V3 Technical Report, 24.12

[2] DeepSeek-AI, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2501

[3] Llama Team, AI @ Meta, The Llama 3 Herd of Models, 24.07

■ 참고 자료

[1] Meta, The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation

[2] SK쉴더스 EQST 그룹, LLM Application 취약점 진단 가이드