AI LLM 진단 절차와 취약점 대응방안 보고서

1. **AI LLM 개요**

   LLM(Large Language Model)은 대규모 텍스트 및 코드 데이터셋을 기반으로 학습된 인공지능 모델로, 자연어 처리(NLP) 분야에서 혁신적인 성과를 보여주고 있다. 대표적으로 GPT, PaLM, Claude 등이 있으며 이들은 텍스트 생성, 번역, 요약, 질의응답, 코드 작성 등 다양한 작업을 수행할 수 있다.

   LLM은 의료, 금융, 교육, 국방 등 다양한 산업에 적용되며 특히 챗봇, 검색 증강 생성(RAG), 자동화된 고객 응대 시스템 등에서 활용도가 높다.

2. **LLM 주요 취약점**

   LLM은 강력한 기능만큼 다양한 보안 취약점을 내포하고 있다. OWASP와 주요 보안 기관들은 다음과 같은 취약점을 경고하고 있다:

취약점 유형	설명
프롬프트 인젝션	악의적 입력을 통해 모델의 의도된 동작을 우회하거나 조작
민감 정보 노출	학습 데이터 또는 출력에서 개인정보나 기업 기밀이 유출
공급망 취약점	외부 모델, 플러그인, 학습 데이터의 보안 결함
데이터 및 모델 오염	학습 데이터에 백도어, 편향, 악성 정보 삽입
부적절한 출력 처리	모델이 생성한 출력이 검증되지 않고 시스템에 전달됨
과도한 자율성	모델이 권한 이상의 작업을 수행하거나 시스템을 제어
시스템 프롬프트 유출	내부 지침이나 설정이 외부에 노출되어 악용 가능
벡터 및 임베딩 취약점	검색 증강 생성(RAG)에서 벡터 DB 조작 가능성

3. **LLM 취약점 대응 방안**

   LLM의 보안 위협에 대응하기 위해 다음과 같은 절차와 전략이 필요하다:

   - 진단 절차
     - 사전 협의 및 범위 선정: 모델 정보, 학습 데이터, 소스 코드, 배포 환경 확인
     - 대상 분석 및 계획 수립: 서비스 구조, 데이터 흐름, 테스트 시나리오 준비
     - 위협 분석: 예상 공격 시나리오 수립 및 모델 특성 분석
     - 취약점 점검 및 대응 방안 수립: 실제 공격 시뮬레이션 및 개선 제안
   - 대응 전략
     - 프롬프트 엔지니어링: 시스템 프롬프트에 명확한 역할과 제한 명시
     - 입출력 검증: 입력 및 출력에 대한 필터링과 검증 절차 마련
     - 휴먼 인 더 루프(HITL): 민감 작업에 인간의 승인 절차 포함

- 샌드박싱 및 격리: 외부 코드 실행 시 시스템 자원 보호
- 공급망 검증: 외부 모델 및 데이터의 신뢰성 확인
- 취약점 스캐닝 및 패치 관리: 구성 요소의 지속적인 보안 점검

4. **AI LLM에 대한 종합적인 고찰**

LLM은 기술적으로 놀라운 진보를 이루었지만 그만큼 위험도 함께 커지고 있다. 특히 프롬프트 인젝션은 단순한 입력만으로도 모델을 조작할 수 있어 매우 치명적이다. 이는 기존 보안 개념과는 다른 새로운 위협이며 보안 담당자들이 기존의 웹 보안 프레임워크만으로는 대응하기 어렵다.

또한 LLM은 인간처럼 자연어를 이해하고 생성하지만, 그 판단은 통계적 확률에 기반한다. 따라서 “신뢰”라는 개념을 적용하기 어렵고 민감한 업무에 투입될 경우 반드시 인간의 감독이 필요하다.

결국 LLM의 도입은 기술적 편의성과 보안 리스크 사이의 균형을 요구한다. 기업과 기관은 단순히 기능 구현에 집중하기보다는 보안과 윤리적 책임을 함께 고려한 전략을 수립해야 한다. LLM은 도구이지 판단 주체가 아니며 인간의 통제 아래에서만 그 진가를 발휘할 수 있다.