2025년도 캡스톤 디자인 1 종합설계 프로젝트 요약 계획서

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **팀명** | VIGILANT | | **제출일** | | 2025년 03월 20일 | |
| **지도교수** | 윤명근 교수 | | | | | |
| **프로젝트제목** | 내부 정보 유출 위험 감지를 위한 생성형 AI 보안 기술 | | | | | |
| **팀원** | **성명** | **학번** | | **e-mail 주소** | | **SNS ID** |
| 조항범 (조장) | 20203145 | | hururuk@kookmin.ac.kr | | always-spring01 |
| 오주엽 | 20222120 | | to765@kookmin.ac.kr | | JuYeobOh |
| 박찬민 | 20223173 | | pbkpcm@kookmin.ac.kr | | chanmin7250 |
| 백승렬 | 20202086 | | imvst4@gmail.com | | komi22 |
| 심지아 | 20213022 | | namu2054@kookmin.ac.kr | | za1014 |
| 김민승 | 20203038 | | brobb@kookmin.ac.kr | | MinSeugn |
| **설계**  **프로젝트**  **개요** | **목표**: 본 프로젝트는 기업 및 기관이 생성형 인공지능 서비스를 안전하게 활용할 수 있도록, 외부 AI 서비스에 전달되는 사용자 질의를 사전에 점검하고 민감 정보 유출을 방지하는 중간 필터링 및 제어 시스템을 개발하는 것을 목표로 한다.  **개발 내용**: 사용자의 질의를 벡터 임베딩으로 변환하여 기밀 문서와의 의미적 유사도를 분석하는 탐지 기능을 개발한다. 또한, RAG(Retrieval-Augmented Generation) 기반 내부 LLM을 적용하여 외부 AI 서비스 차단 없이 안전한 응답을 제공하는 시스템을 구축한다. 더불어, Streamlit 기반 관리자 대시보드를 통해 기밀 문서 업로드, 관리 및 실시간 탐지 로그 확인 기능을 제공한다.  **결과물 내역**: 사용자 GUI에서는 ChatGPT API 또는 내부 AI 모델을 활용하여 질의의 기밀 문서 포함 여부를 실시간 분석한다. 관리자 GUI에서는 기밀 문서 업로드, 실시간 탐지 기록 확인 및 로그 관리 기능을 제공하여 정보 보호 체계를 강화한다.  **현실적 제한요소**: 형식적 유사도 비교는 명확하지만, 문맥적 유사도 판단 기준이 표준화되지 않아 탐지의 주관성이 존재한다. 또한, 대용량 기밀 문서 처리 시 연산량이 증가하여 유사도 비교 속도에 제한이 발생할 수 있다.  **관련 현황 및 배경**: ChatGPT, Bard, Claude 등 다양한 생성형 AI 서비스가 기업 내 업무 자동화, 문서 작성, 코드 개발, 데이터 분석 등의 분야에서 적극적으로 활용되고 있다. 그러나 이러한 AI 서비스에 업무 관련 질의를 입력하는 과정에서 의도치 않게 기업 내부의 기밀 정보가 유출되는 사례가 빈번히 발생하고 있으며, 이에 대한 보안 대책 마련이 필수적이다.  **차별화 전략**: 기존 키워드 검색을 넘어 문맥적 유사도를 분석하여 기밀 유출을 보다 정교하게 방어한다. 또한, 일관된 탐지 알고리즘 및 통계 제공을 통해 지속적인 보안 강화가 가능하도록 지원한다. 외부 AI 차단이 아닌 내부 LLM을 활용하여 생산성을 유지하면서도 안전한 AI 응답을 제공하는 것도 주요 차별점이다.  **기대효과**: 본 프로젝트를 통해 기업의 기밀 정보 유출을 사전에 방지하고, AI를 안전하게 활용할 수 있는 환경을 구축한다. 이를 통해 기업은 보안성을 강화하면서도 AI 서비스를 신뢰할 수 있는 방식으로 도입하고 운영할 수 있다. | | | | | |

(주의)요약제안서는 1페이지로만 작성한다.(폰트크기 변경불가, 1페이지이상 작성시 제출불가)