2025년도 캡스톤 디자인 1 종합설계 프로젝트 요약 계획서

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **팀명** |  | | **제출일** | | 2025년 03월 20일 | |
| **지도교수** | 윤명근 교수 | | | | | |
| **프로젝트제목** | 안전한 생성형 인공지능 사용 기술 개발 | | | | | |
| **팀원** | **성명** | **학번** | | **e-mail 주소** | | **SNS ID** |
| 조항범 (조장) | 20203145 | | [hururuk@kookmin.ac.kr](mailto:hururuk@kookmin.ac.kr) | | always-spring01 |
| 오주엽 | 20222120 | |  | |  |
| 박찬민 | 20223173 | |  | |  |
| 백승렬 | 20202086 | |  | |  |
| 심지아 | 20213022 | |  | |  |
| 김민승 | 20203038 | |  | |  |
| **설계**  **프로젝트**  **개요** | **목표:** 본 프로젝트의 목표는 기업 및 기관에서 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 것이다.  **개발 내용:** 기밀 문서와 의미적으로 유사한 질의문, 질의문서를 탐지하는 기능을 개발한다. 이를 위해 사용자의 질의를 임베딩하여 벡터로 변환한 후, 사전에 저장된 기밀 문서와의 유사도 측정 기능을 개발한다. 또한 RAG 기반 내부 LLM 모델을 통한 답변 기능을 개발한다. Streamlit 기반 관리자 페이지를 구현하여 업로드된 기밀 문서 관리 혹은 실시간 탐지 로그를 시각적으로 확인할 수 있는 기능을 제공한다.  **결과물 내역:** 사용자 GUI 및 ChatGPT API를 통해 시연자가 직접 기밀문서 관련 질의문을 작성해보고 해당 질의문이 기밀문서에 포함된 내용인지 탐지 결과를 확인할 수 있다. 또한 관리자 GUI를 통해 기밀문서 관리가 가능하다. 사용자는 기밀 문서를 등록할 수 있고, 실시간으로 기밀 유출 탐지 모니터링이 가능하다.  **현실적 제한요소:** 형식적 유사도를 비교하는 기준은 명확하게 정의 가능하지만, 문맥적 유사도를 판단하는 기준이 표준으로 정의되어 있지 않다는 제한 요소가 있다. 또한 서버에 저장된 기밀 문서의 종류가 많고, 각 문서의 크기가 클수록 연산량이 많아지고 유사도 비교 과정에 많은 시간적 제한 요소가 생긴다.  **관련 현황 및 배경:** ChatGPT, Bard, Claude 등 다양한 생성형 인공지능 서비스가 등장하며 기업 내에서 업무 자동화, 보고서 작성, 코드 개발, 데이터 분석 등 다양한 시도가 활발히 이루어지고 있다. 그러나 업무상 생성형 AI에 질의하는 과정에서 의도치 않게 내부 기밀 정보를 유출하는 사례가 빈번히 발생하고 있다.  **차별화 전략:** 본 프로젝트의 차별화 전략은 문맥간 유사도 비교, 일관적인 알고리즘, 기밀 유출 통계 제공에 있다. 기존의 형식적 유사도의 한계에서 벗어나 문맥적으로 유사한 질의를 탐지하여 기업 내 기밀 유출을 효과적으로 방어할 수 있다. 일관적인 알고리즘, 통계 자료 제공을 통해 안정화된 서비스를 제공하는 것 또한 차별화 전략 중 하나이다.  **기대효과:** 본 프로젝트를 통해 기업 내부의 기밀 정보 유출을 방지하며 인공지능 기술을 활용할 수 있다. | | | | | |

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