안전 가드레일 (Safety Guardrails) 구현

AI 시스템이 안전한 범위 내에서만 작동하도록 보장하는 제약 조건

입력 가드레일

- 허용된 입력 범위 제한
- 악의적 프롬프트 탐지
- 개인정보 자동 제거
- 형식 검증

출력 가드레일

- 해로운 내용 필터링
- 의학적 타당성 검증
- 불확실성 명시
- 면책 조항 추가

행동 가드레일

- 허용된 작업만 실행
- 권한 기반 접근 제어
- 감사 로그 자동 기록
- 임계값 기반 경고

동적 가드레일

- 컨텍스트 기반 조정
- 사용자 역할별 제한
- 실시간 위험 평가
- 적응형 임계값