피싱과 개인정보 유출을 막는 '양방향 보안' 에이전트 [요약]

□ 제안배경

○ **(생활의 중심이된 카카오톡)** 카카오톡은 일상의 대화뿐만 아니라 신분증, 여권, 카드번호, 비밀번호 등 민감한 개인정보를 주고받는 핵심 채널

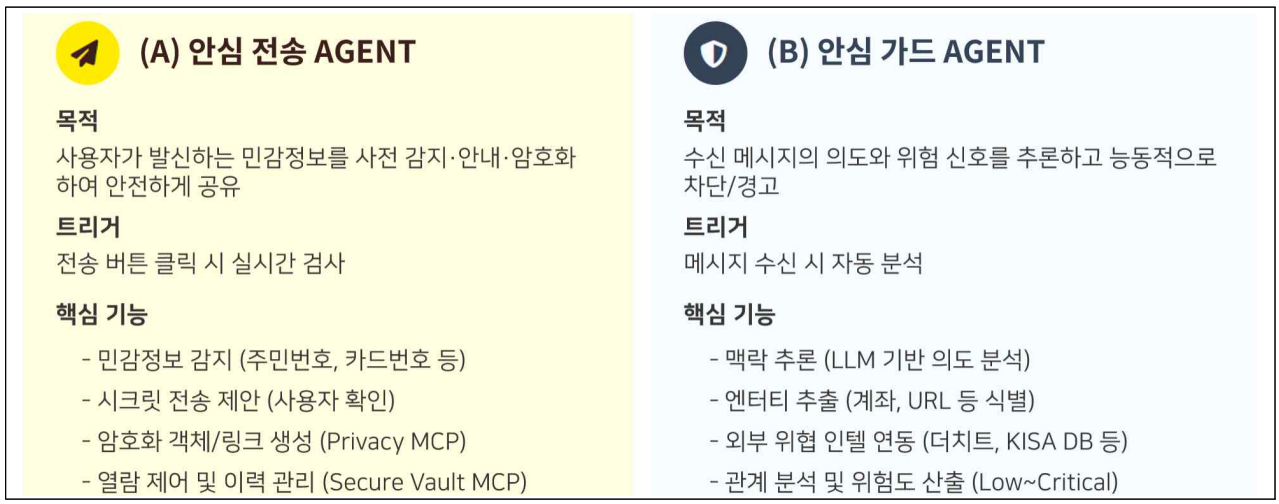
○ **(문제 ①) 발신 위험 : '의도적 전송'으로 인한 민감정보 유출**

- 피싱 위협만큼이나 심각한 문제는 사용자가 '필요에 의해(의도적으로)' 민감정보를 전송할 때 발생하는 '발생 위험'이 있음

○ **(문제 ②) 수신 위협 : 지능화된 '메신저 피싱'의 폭증**

- 금융 범죄가 단순 금전 요구를 넘어서 "엄마, 폰 고장났어", "급전 필요해" 등 맥락을 이용한 방법을 넘어서 최신 피싱 공격은 더욱 진화

□ 개선방안 : (A)안심 전송 Agent + (B)안심 가드 Agent



□ 우수성

○ **(의도기반 능동형 Agent)** 의도와 맥락을 추론하는 Agentic AI

○ **(프라이버시)** 하이브리드 AI 기반의 선제적 정보보호

○ **(양방향 보안)** 메시지 피싱 방어, 개인정보 보호를 에이전트가 통합관리

□ 기대효과

○ **(사용자)** 민감정보 전송 시 '보낸 후 불안감'을 해소하고 피해 사전예방

○ **(플랫폼/사회)** 사회 안전망 구축 및 범죄 예방

○ **(비즈니스)** 신뢰 기반의 B2B 신규 비즈니스 확장

피싱과 개인정보 유출을 막는 '양방향 보안' 에이전트

□ 제안배경 : 왜 양방향 보안이 필요한가?

○ **(생활의 중심이된 카카오톡)** 카카오톡은 일상의 대화뿐만 아니라 신분증, 여권, 카드번호, 비밀번호 등 민감한 개인정보를 주고받는 핵심 채널

○ **(문제 ①) 발신 위험 : '의도적 전송'으로 인한 민감정보 유출**

- 피싱 위협만큼이나 심각한 문제는 사용자가 '필요에 의해(의도적으로)' 민감정보를 전송할 때 발생하는 '발생 위험'이 있음
  - · 사용자가 가족에게 인증을 위해 신분증 사진을 보내거나, 결제를 위해 카드 사진을 보낼 때, 카카오톡 '메시지 삭제' 기능은 사후 대응에 불과
  - · 수신자가 해당 메시지를 '캡처'하거나 '저장'하는 순간, 데이터는 발신자의 통제권을 벗어나 2차 유출 위험에 무방비로 노출
  - · 메신저를 통한 개인정보 유출은 플랫폼의 신뢰도와 직결되는 문제*

* 개인정보보호위원회는 메신저 오픈채팅방의 취약점을 이용한 개인정보 유출 사례에 대해 사업자에게 막대한 과징금을 부과하는 등 플랫폼이 사용자의 '발신 위험'까지 관리해야 할 책임이 있음을 시사함

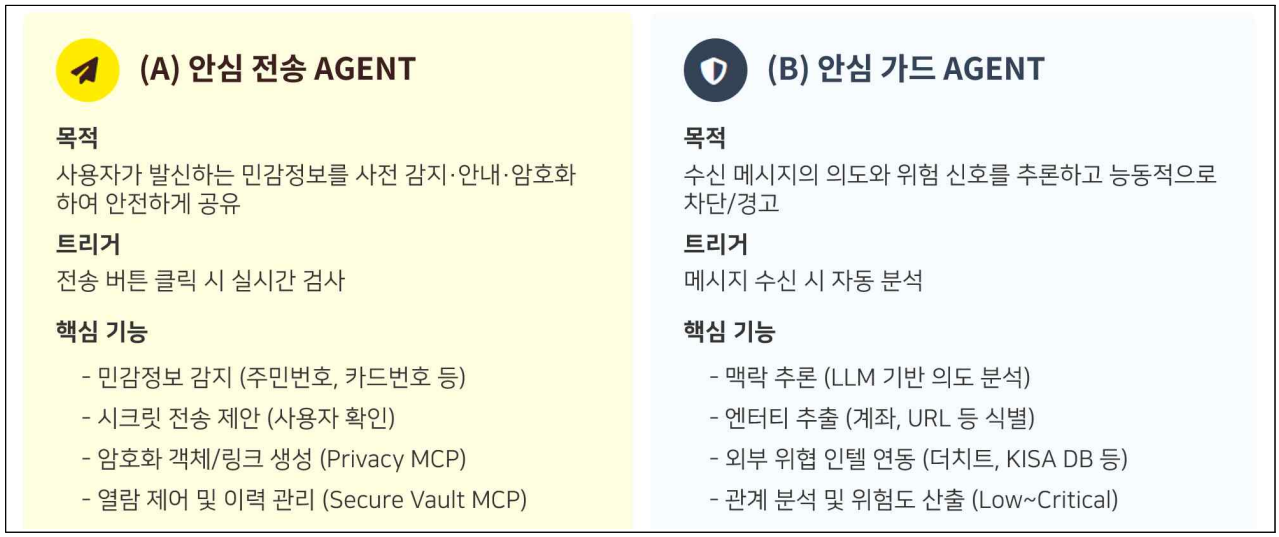
○ **(문제 ②) 수신 위협 : 지능화된 '메신저 피싱'의 폭증**

- 금융 범죄가 단순 금전 요구를 넘어서 "엄마, 폰 고장났어", "급전 필요해" 등 맥락을 이용한 방법을 넘어서 최신 피싱 공격은 더욱 진화
  - · 2025년 1분기 보이스피싱 피해액은 3,116억원에 달하며, 이는 전년 동기 대비 2.2배 급증
  - · KISA는 2025년 사이버 위협 전망 1순위로 공격자의 생성형AI 활용 본격화를 제시하는 등 現키워드 기반의 필터링은 피싱 방지에 한계
  - · 실제로 AI가 생성한 피싱 메시지의 경우 문법적 오류가 없고, 개인화된 맥락을 제공하며, 고도의 사회공학적 접근으로 더욱 치밀해 짐

○ **(목표) 양방향 통합 보안 Agent 구축**

- 메신저에서 '보낼 때의 불안감'과 '받을 때의 위협'을 AI 에이전트가 능동적으로 해결하는 '양방향 통합 보안' Agent를 구축

□ 개선방안 : (A)안심 전송 Agent + (B)안심 가드 Agent



(A) 안심 전송 AGENT 동작 흐름

※ 사용자가 민감정보를 보낼 때

① 감지 (Kanana)

- 사용자가 전송 버튼 클릭 시 텍스트/이미지(OCR) 실시간 분석
- 주민번호, 카드번호, 여권 등 민감정보 패턴 감지

② 제안 (AI 능동적 개입)

- '민감정보가 감지되었습니다. 시크릿 전송하시겠습니까?'
- 사용자에게 선택할 수 있도록 함

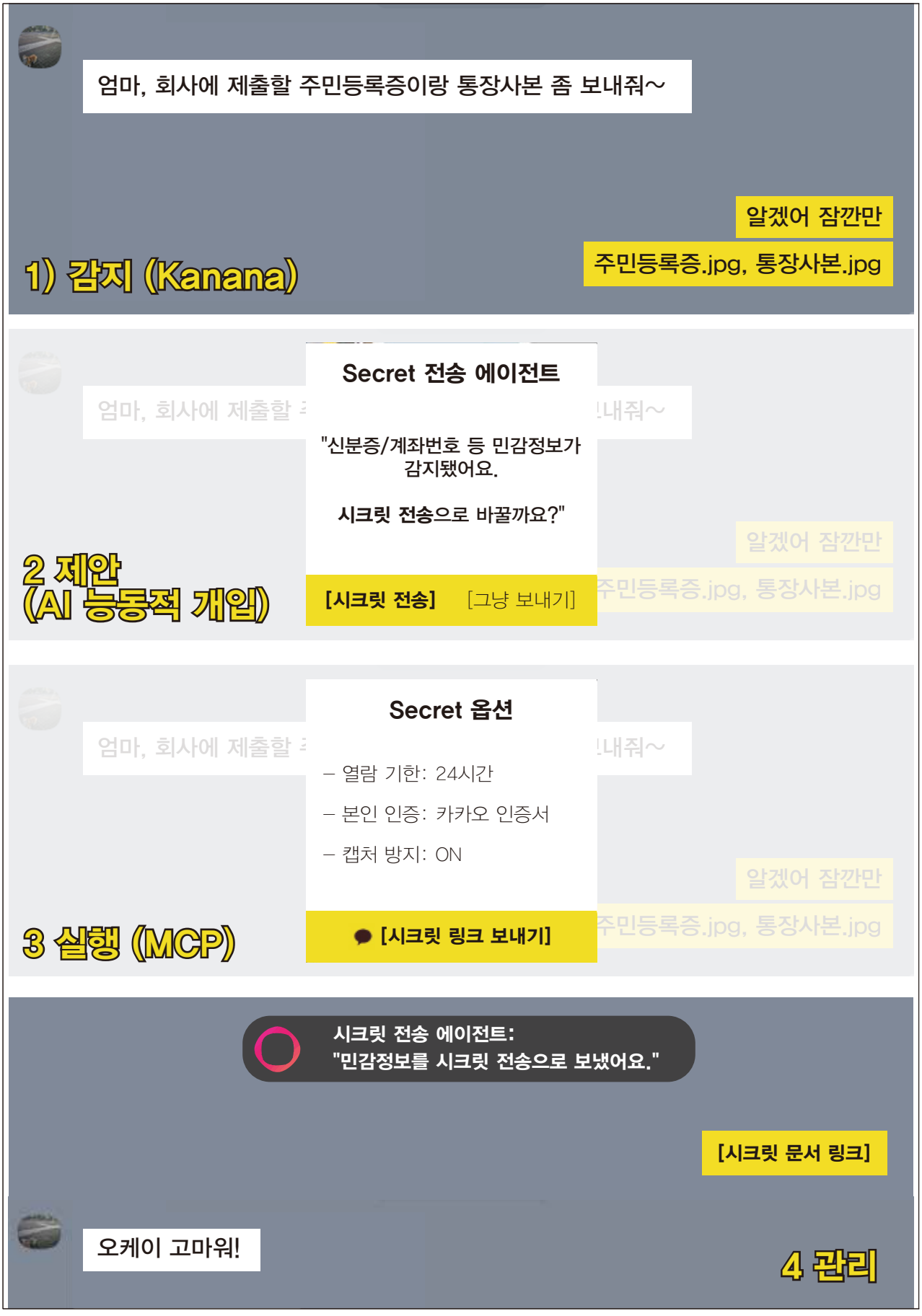
③ 실행 (MCP)

- Privacy MCP : 민감정보 탐지
- Secure Vault MCP : 암호화 객체 생성 및 링크 전송
  - · 열람 기한 설정 (1시간 / 24시간 / 7일)
  - · 수신자 본인인증 (카카오인증 / 생체인증)
  - · 캡처 방지 뷰어
- Verification MCP : 카카오 인증 API 연동

④ 관리

- 전송 후에도 열람 기한 조정, 즉시 삭제 가능
- 열람 이력 실시간 확인

※ 안심 전송 Agent의 사용 예시



(B) 안심 가드 AGENT 동작 흐름 ※ 사용자가 채팅을 수신할 때

① 맥락 분석 (Kanana LLM + Context Analyzer MCP)

- 수신 메시지의 '의도' 추론
  - 돈, 계좌와 같은 키워드 탐지가 아닌 LLM의 추론 능력을 활용하여 문장의 '의도'를 파악 *(예: 가족사칭, 긴급상황연출, 금전요구 등)*
- 종합적 맥락 이해를 기반으로 위험도 점수화

② 정보 추출 (Entity Extractor MCP)

- 메시지 내 정보 식별
  - 계좌번호, 전화번호, URL, 카드번호 등
- NER(Named Entity Recognition) 기술 활용

③ 교차 검증 (Threat Intelligence MCP)

- 외부 보안 API 실시간 연동
  - 더치트(계좌번호 사기), KISA 피싱 신고 DB, Google Safe Browsing
- 식별된 정보의 블랙리스트 여부 확인

④ 관계 분석 (Social Graph MCP)

- 카카오톡 대화 히스토리 분석
  - 처음 대화하는 사람인지 판단
  - 기존 관계 신뢰도 평가 (대화하던 사람과 유사한지 검증)

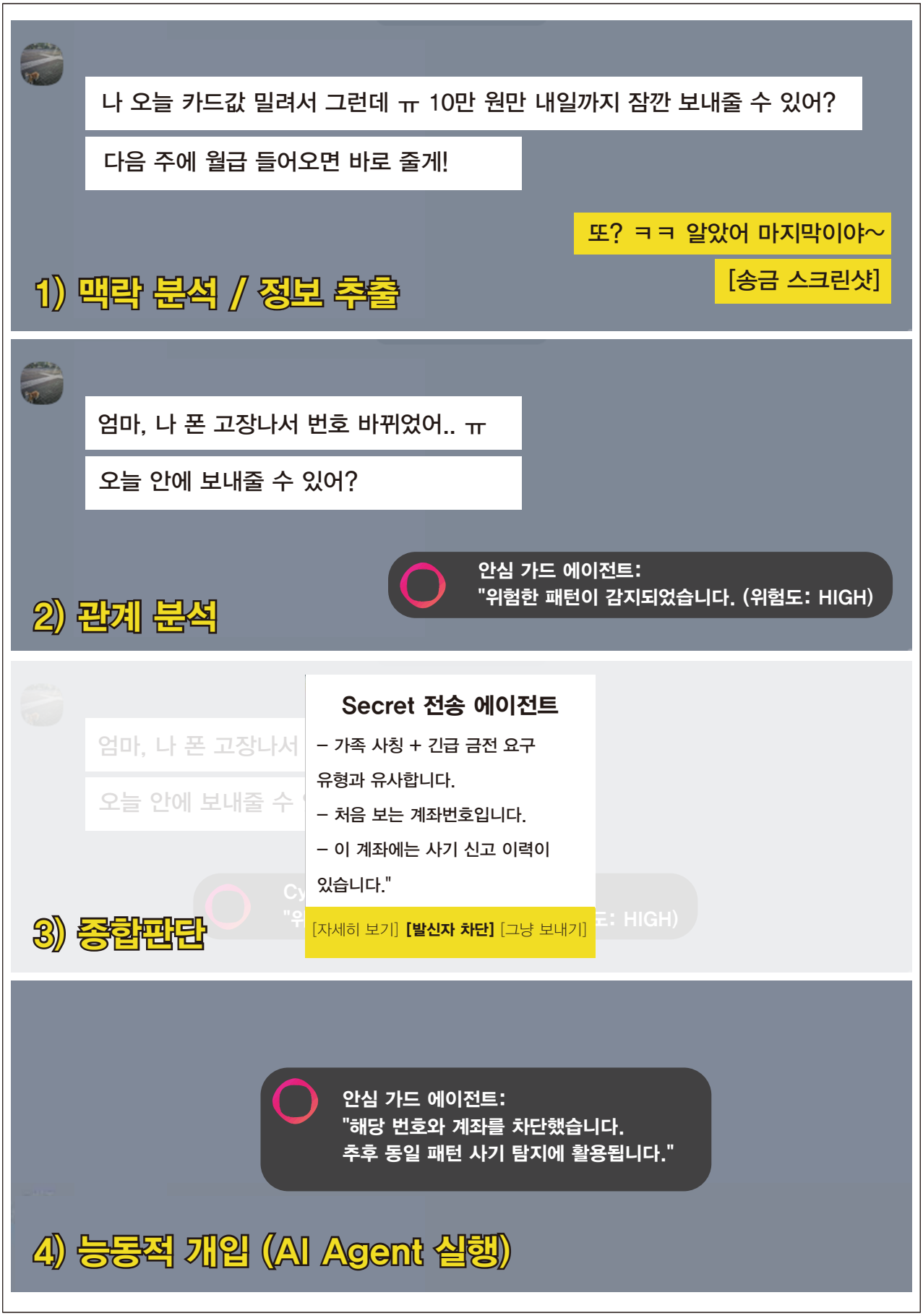
⑤ 종합 판단 (Decision Engine MCP)

- 1~4단계 결과 통합 분석
- 최종 위험도 산출 (Low / Medium / High / Critical)

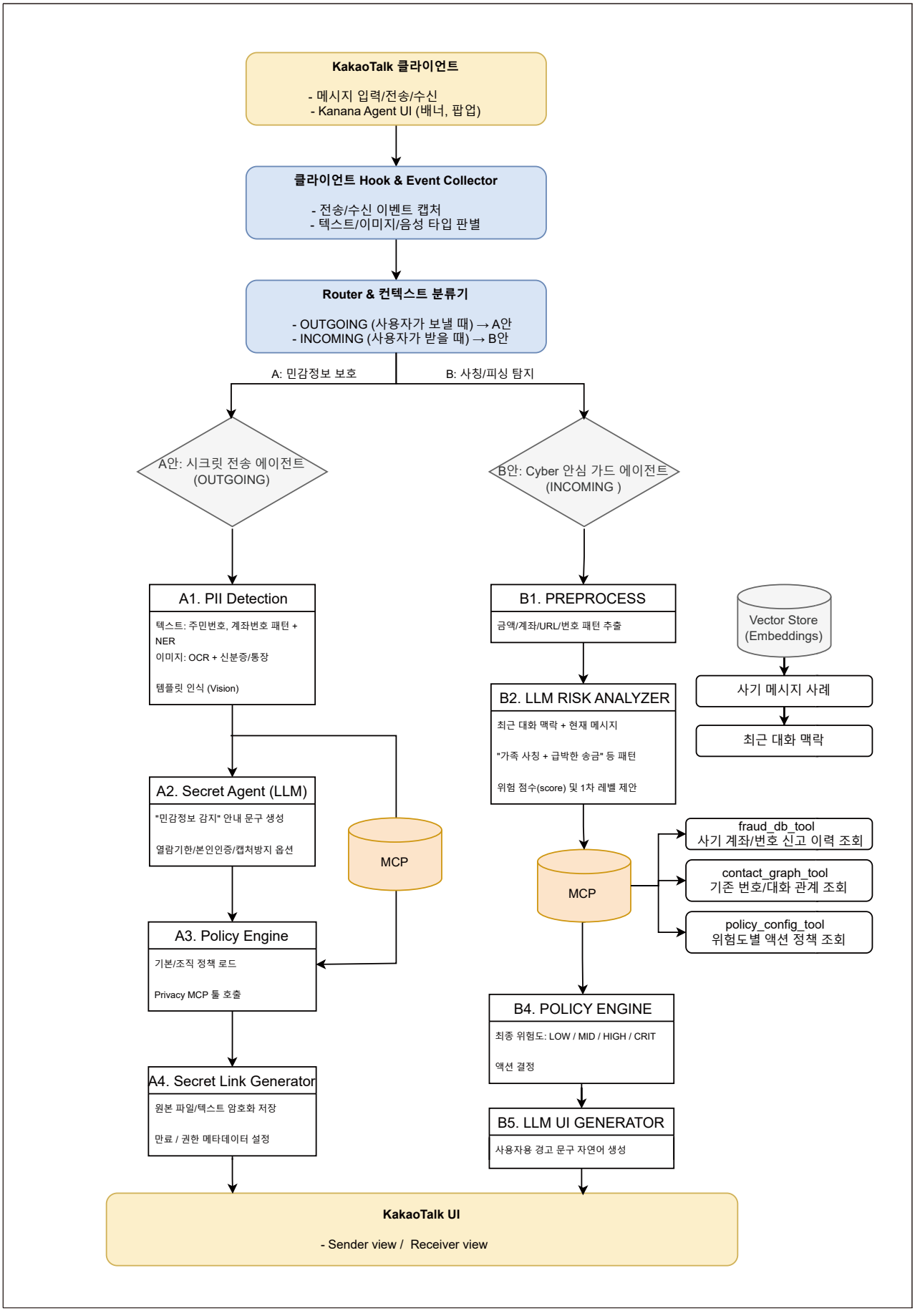
⑥ 능동적 개입 (AI Agent 실행)

- (Low) 정보성 알림 *(예: 처음 보는 계좌입니다)*
- (Medium) 주의 경고 *(예: 유사 사기 사례 신고됨)*
- (High) 강력 차단 *(예: 사기 신고 5건 접수된 계좌)*
- (Critical) 즉시 차단 + 경찰청 연계

※ 안심 가드 Agent의 사용 예시



□ 아키텍처


KakaoTalk 클라이언트
- 메시지 입력/전송/수신
- Kanana Agent UI (배너, 팝업)
클라이언트 Hook & Event Collector
- 전송/수신 이벤트 캡처
- 텍스트/이미지/음성 타입 판별
Router & 컨텍스트 분류기
- OUTGOING (사용자가 보낼 때) → A안
- INCOMING (사용자가 받을 때) → B안
A: 민감정보 보호
B: 사칭/피싱 탐지
A안: 시크릿 전송 에이전트 (OUTGOING)
B안: Cyber 안심 가드 에이전트 (INCOMING)
A1. PII Detection
텍스트: 주민번호, 계좌번호 패턴 + NER
이미지: OCR + 신분증/통장
템플릿 인식 (Vision)
A2. Secret Agent (LLM)
"민감정보 감지" 안내 문구 생성
열람기한/본인인증/캡처방지 옵션
MCP
A3. Policy Engine
기본/조직 정책 로드
Privacy MCP 툴 호출
A4. Secret Link Generator
원본 파일/텍스트 암호화 저장
만료 / 권한 메타데이터 설정
B1. PREPROCESS
금액/계좌/URL/번호 패턴 추출
B2. LLM RISK ANALYZER
최근 대화 맥락 + 현재 메시지
"가족 사칭 + 급박한 송금" 등 패턴
위험 점수(score) 및 1차 레벨 제안
Vector Store (Embeddings)
사기 메시지 사례
최근 대화 맥락
MCP
fraud_db_tool
사기 계좌/번호 신고 이력 조회
contact_graph_tool
기존 번호/대화 관계 조회
policy_config_tool
위험도별 액션 정책 조회
B4. POLICY ENGINE
최종 위험도: LOW / MID / HIGH / CRIT
액션 결정
B5. LLM UI GENERATOR
사용자용 경고 문구 자연어 생성
KakaoTalk UI
- Sender view / Receiver view

□ 기술 스펙 (모델, 프레임워크/툴, MCP/API 등)

구분	명칭	주요 내용	비고
LLM	Kanana	카카오톡 대화 맥락 이해, 위험도 분석, 경고가이드 문장 생성	외부 활용
Vision	Vision/OCR 모델	신분증 · 통장 등 문서 이미지 인식, OCR 기반 PII 후보 영역 검출	외부 활용
MCP	Privacy MCP	민감정보(주민번호, 계좌번호 등) 탐지 정책 및 PII 타입 관리	개발
MCP	Verification MCP	카카오 인증/본인인증 API 연동, 민감 정보 열람 시 추가 인증 수행	개발
MCP	Entity Extractor MCP	카카오톡 대화 히스토리 · 연락처 기반 관계/신뢰도 분석	개발
프레임워크	FastAPI or Django	Kanana 백엔드 REST API 서버 구현, MCP 엔드포인트 제공	개발
프레임워크	Docker	MCP 서비스 및 백엔드 컨테이너화, 배포 관리	인프라
TOOL	Prometheus / Grafana	MCP · 백엔드 서비스의 성능/상태 모니터링 대시보드	운영
TOOL	ELK Stack	메시지 · 위험 탐지 로그 수집 및 검색/분석	운영
API	Kakao 메시지 이벤트 API	카카오톡 메시지 전송/수신 이벤트 웹훅 연동	외부 활용
API	Kakao 인증 API	카카오 인증서/로그인 기반 사용자인증	외부 활용

□ 우수성

○ **(의도기반 능동형 Agent)** 의도와 맥락을 추론하는 Agentic AI

- 단순 키워드 필터링이 아닌, Kanana LLM의 심층 추론을 통해 “가족 사칭”, “긴급 상황 연출” 등 사회공학적 의도 자체를 파악
- Agent 스스로 ‘계획’을 수립하고, MCP와 Agentic RAG를 자율적으로 실행하여 변종 피싱 공격까지 능동적으로 방어

○ **(프라이버시)** 하이브리드 AI 기반의 선제적 정보보호

- 안심 전송 Agent 동작 시 온디바이스 AI Kanana가 사용자 기기내에서 민감정보를 처리하고, 고성능 추론만 서버로 처리

○ **(양방향 보안)** 메시지 수신 시 피싱 방어, 발신 시 개인정보 보호를 단일 에이전트가 통합하여 관리

□ 기대효과

○ **(사용자)** 민감정보 전송 시 ‘보낸 후 불안감’을 해소하고 피해 사전예방

- 사용자는 어떠한 상황에서도 카카오톡을 안심하고 사용할 수 있으며 이는 플랫폼에 대한 신뢰도 향상으로 이어짐

○ **(플랫폼/사회)** 사회 안전망 구축 및 범죄 예방

- 디지털 취약계층을 AI 메신저 피싱 범죄로부터 보호하는 강력한 ‘사회 안전망’ 역할을 수행하며, 피싱 범죄 피해액을 절감할 수 있음

○ **(비즈니스)** 신뢰 기반의 B2B 신규 비즈니스 확장

- 비대면 계약서, 법무/의료 상담 등 고도의 보안이 요구되는 전문 영역으로 카카오톡 활용성 확장