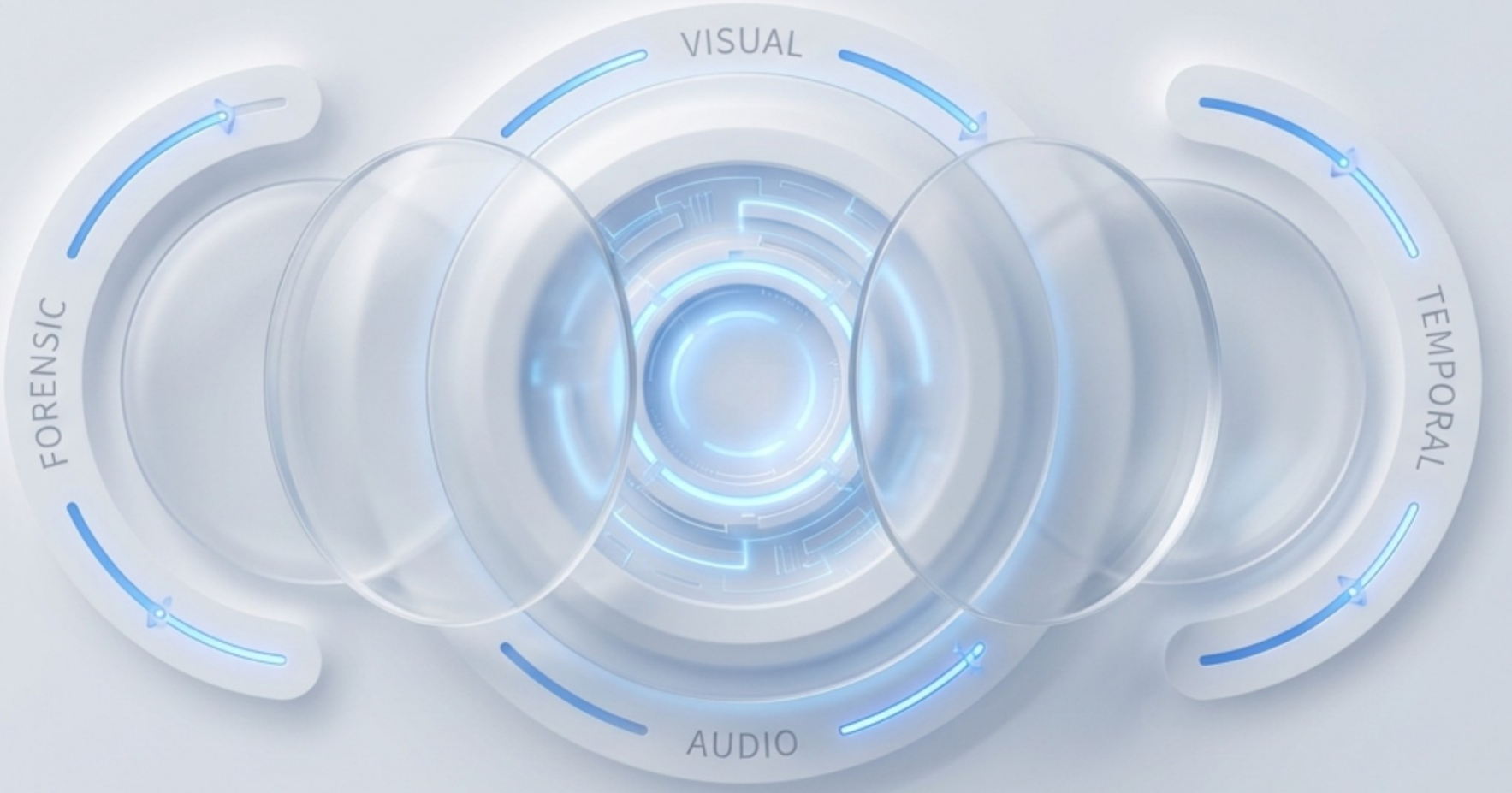
TruthLens v4.4: 차세대 멀티모달 딥페이크 탐지 플랫폼

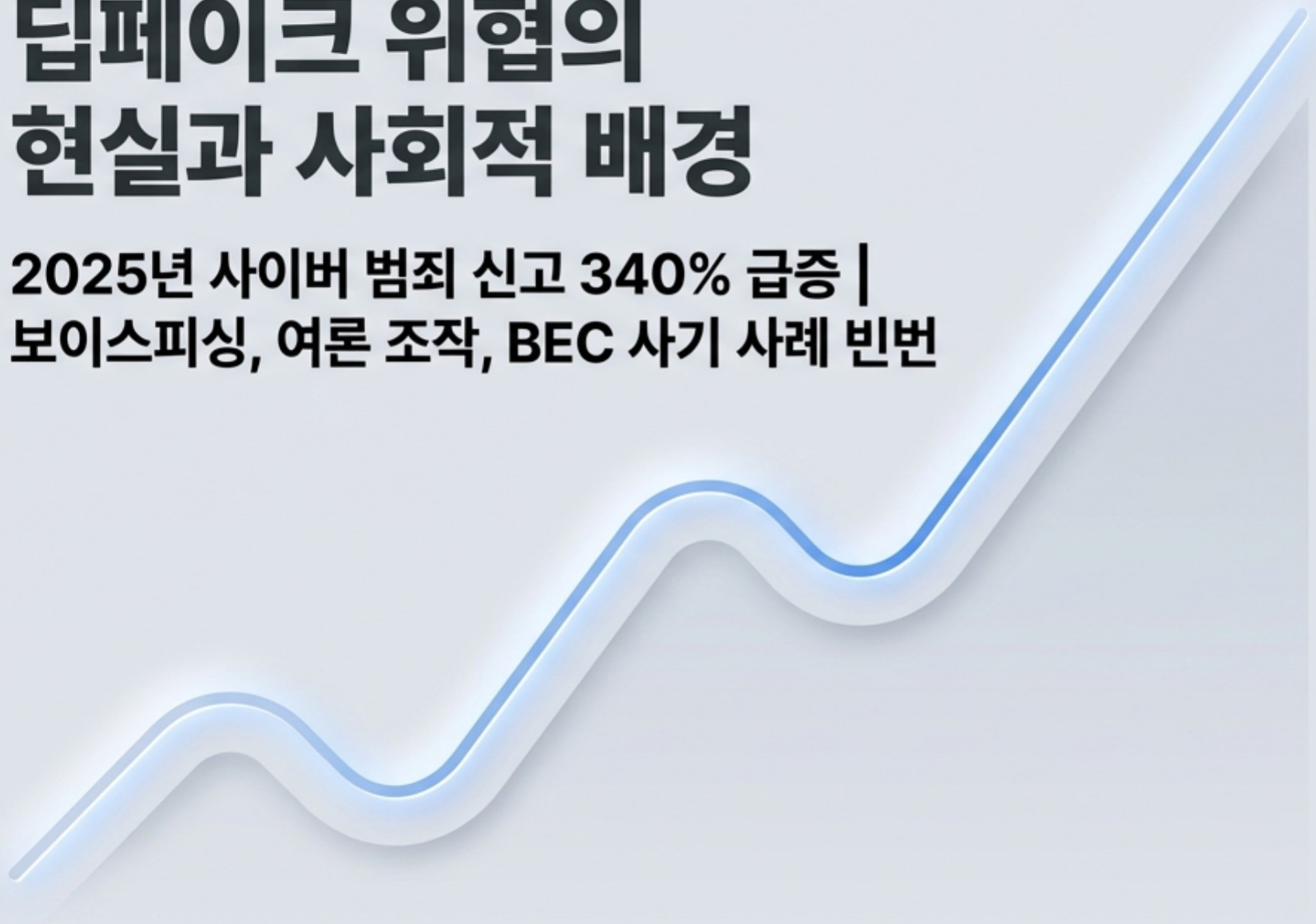
40개 탐지 모듈과 멀티 에이전트 AI가 결합된 설명 가능한 AI(XAI) 방어 체계



2026년 4월 기준 최신 탐지 인프라 적용

딥페이크 위협의 현실과 사회적 배경

2025년 사이버 범죄 신고 340% 급증 | 보이스피싱, 여론 조작, BEC 사기 사례 빈번



금융

기업 CEO 화상 회의 위조로 830억 원(6,200만 달러) 송금 유도, 가족 사칭 보이스피싱

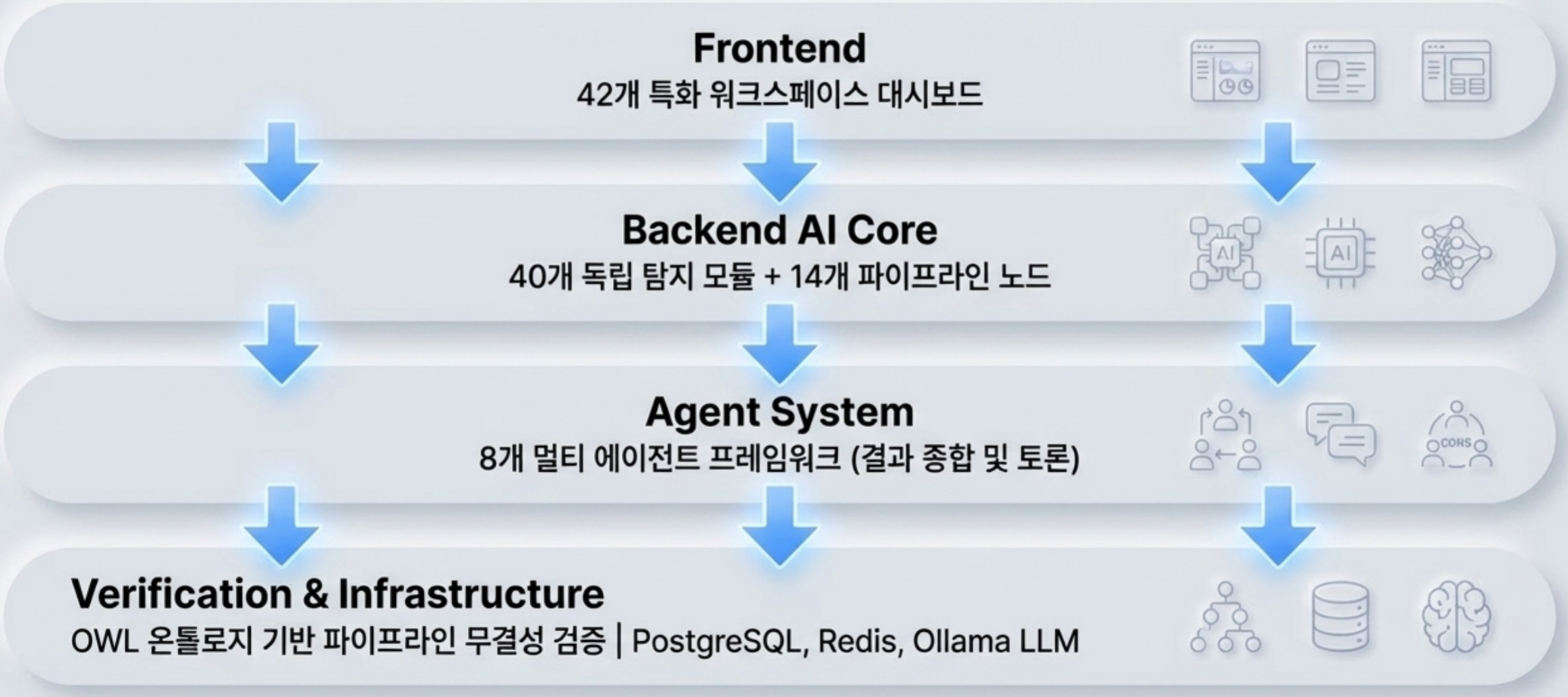
정치/사회

선거 기간 후보자 딥페이크 유포로 민주주의 의사결정 왜곡, 딥페이크 포르노그래피 범죄 확산

방송/미디어

가짜 뉴스 영상 SNS 급증, 유명인 사칭 사기 광고

플랫폼 아키텍처: 다층적 진실 검증 파이프라인



3단계 직관적 판정 시스템과 신뢰도

REAL
(진본)

AI 생성/조작
흔적 미발견

임계값 ≤ 0.25

FAKE
(위조)

AI 생성/조작 흔적
유의미하게 발견

임계값 ≥ 0.65

UNCERTAIN
(판단 유보)

판단 근거 불충분
또는 모호함

*주의: 분석 실패가 아닌
전문가 추가 검토 권고 상태

상황별 맞춤 분석 모드와 사용자 접근성

신속 (Rapid)
10~30초 소요
대량 스크리닝, 긴급 확인

표준 (Standard)
1~5분 소요
일반적인 분석 (권장)

정밀 (Precise)
5~15분 소요
수사 증거, 법적 활용

VIEWER
조회 전용

ANALYST
보고서 및 API 관리

ADMIN
시스템 및 조직 관리

종합 분석 대시보드와 위협 인텔리전스 (BI)

모달리티 방사형 차트

BI 통계 분석

정상
FAKE 비율:
5~15%

임계 위협 경고: 30% 초과 시 유입 경로 긴급 점검

위협 매트릭스

공격 유형 (페이스스왑, 보이스클로닝 등) × 대상 (금융, 정치 등)

[탐지 매트릭스] 딥페이크 탐지를 위한 3대 축

공간/주파수 탐지 (시각)

프레임 단위의 픽셀 조작

확산 모델 노이즈

블렌딩 흔적

시간/생체 분석 (흐름)

프레임 간 일관성

미세 표정

시선 동기화

물리적 포즈 역학

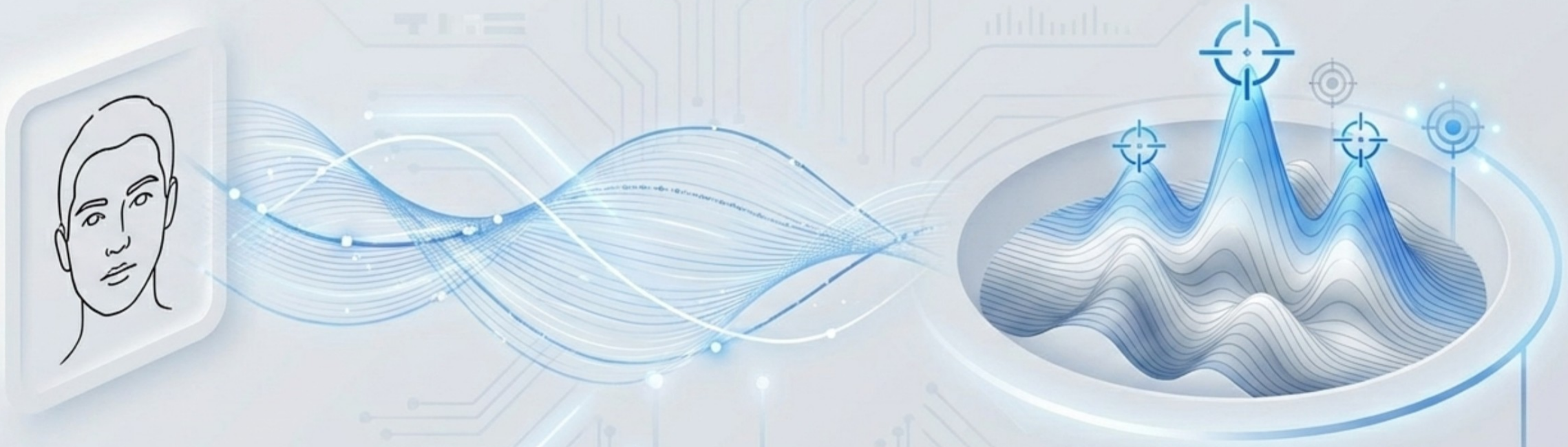
오디오 포렌식 (청각)

AI 음성 합성

립싱크 불일치 밀리초 분석

보코더 흔적

공간 및 주파수 탐지 I: 픽셀 이면의 흔적



DM (확산 모델) 탐지 센터

- Stable Diffusion 등 고유의 재구성 오차(DIRE)를 스크리닝 → 고속 → 정밀 3단계로 분석
- 확률 점수 0.7 이상 시 의심

주파수 분석 (GAN 핑거프린트)

- FFT, DCT, Wavelet 3중 변환 병렬 수행
- GAN 업샘플링 과정의 체커보드 아티팩트 포착

공간 및 주파수 탐지 II: 구조적 불일치 감지

ViT (비전 트랜스포머) 앙상블

- Swin-V2, CrossViT, TIMM 3개 모델 앙상블
- 어텐션 히트맵으로 얼굴 경계, 머리카락-이마, 턱선 주변 식별

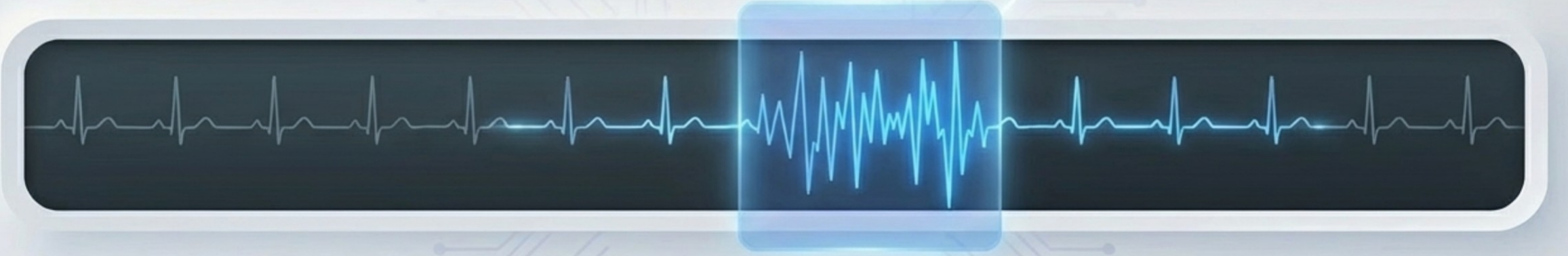
블렌딩 경계 분석

- LAA(학습된 적응적 주의력) 및 푸아송 블렌딩 흔적 역추적
- 교묘하게 조명을 맞춘 고급 딥페이크까지 구조적 불일치로 탐지

시간 및 생체 분석 I: 보이지 않는 부자연스러움

시간 도메인 분석 (Temporal Consistency)

VideoMAE 활용: 특정 구간만 교체된 프레임 스티칭 흔적 급격한 점수 변동 탐지



미세 표정 분석 (Micro-Expression)

FACS 기반 43개 얼굴 근육(AU) 움직임 추적

정상: 분당 2~5회

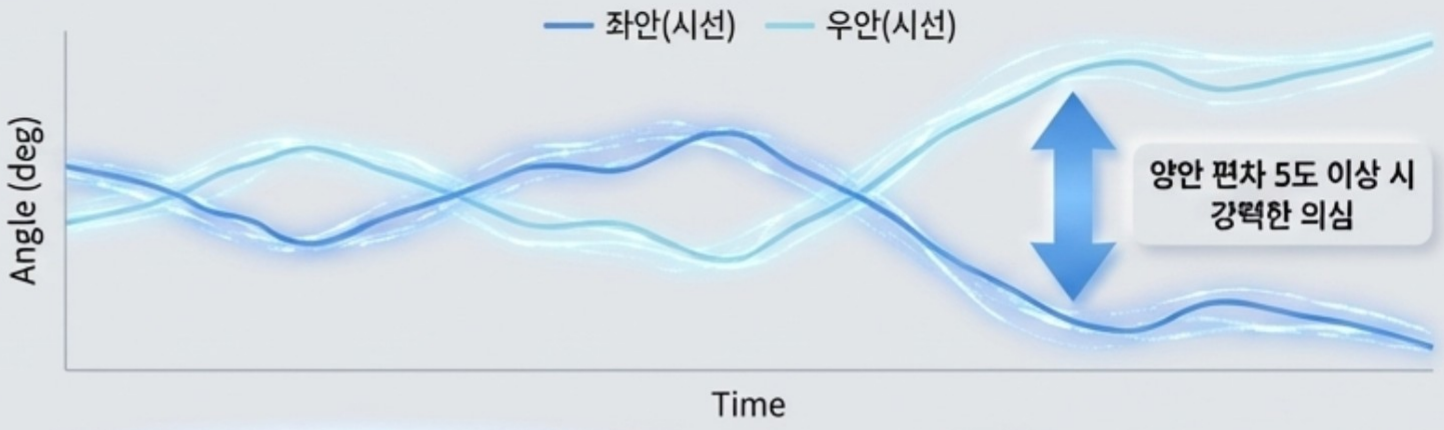
의심: 분당 0~1회 억제

또는 10회 이상 기계적 과장

시간 및 생체 분석 II: 물리 법칙과 시선의 교차 검증

양안(시선) 분석

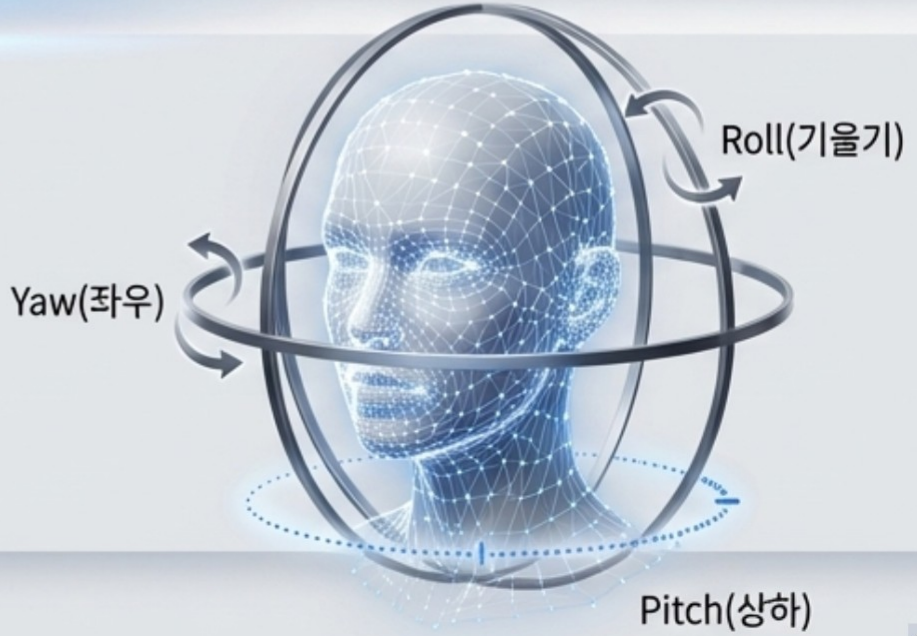
양쪽 눈의 시선 방향 동기화 평가



머리 포즈 역학

Yaw(좌우), Pitch(상하), Roll(기울기)의 물리적 전환 한계점 분석

- 기계적으로 완벽히 매끄러운 곡선은 노이즈가 제거된 조작 증거



오디오 포렌식: AI 음성 합성과 보이스 클로닝 추적

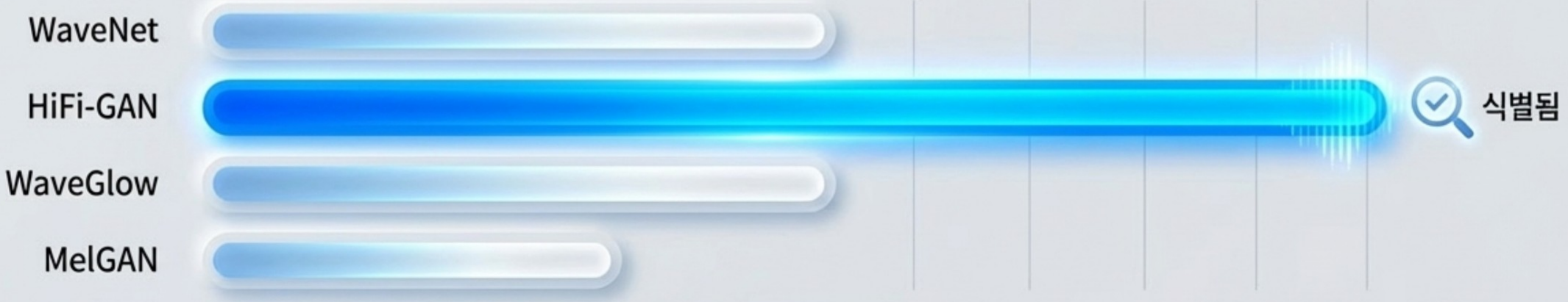
다각적 음성 분석

- MFCC 표준편차 (변동성 결여)
- 스펙트럼 4kHz 이상 단절 탐지

립싱크 포렌식

- 입술 움직임과 음성 동기화 단위 분석
- 지속적 50ms 이상 편차 발생 시 강력한 변조 징후

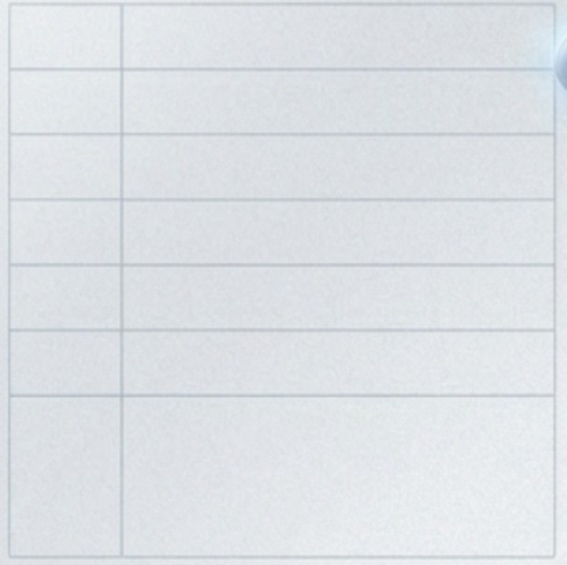
보코더 식별: 20여 종 역추적



포렌식 도구: 법적 증거력 확보와 무결성 증명

포렌식 보고서

수사기관 제출
표준 포맷 자동 생성



해시 체인 검증

SHA-256 해시 체인 및 타임스탬프로 절대적 무결성(VALID) 보장

T-GD 출처 분석과 법적 증거력(LES)

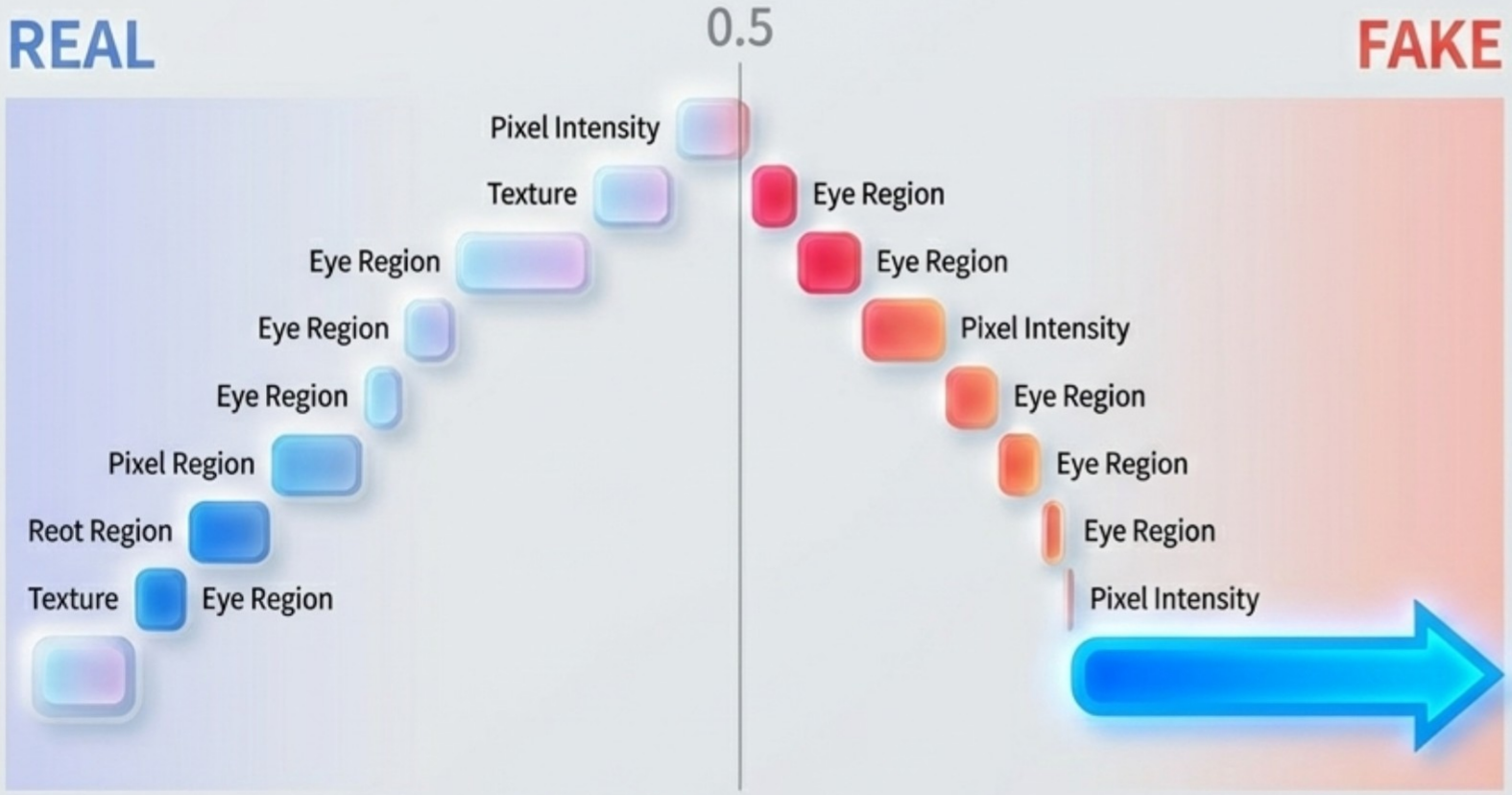
전이학습 기반 생성 탐지 모델
(AUROC 95%+)

수사기관 제출 기준 상희하는
75점 이상의 LES 도출

XAI 심층 분석: 블랙박스를 여는 설명 가능한 AI

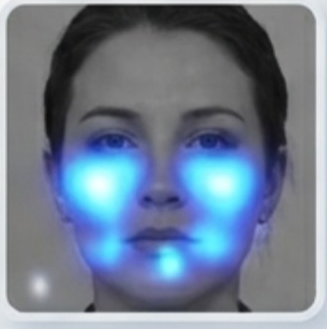
SHAP 폭포(Waterfall) 차트

기저값(0.5)에서 출발하여 40개 모듈이 판정(FAKE/REAL)에 누적 기여한 수치

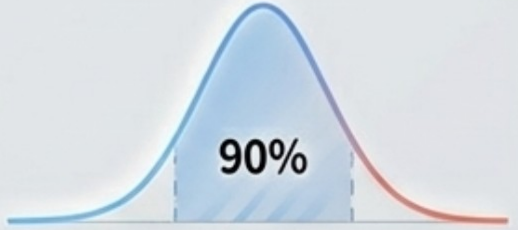


Grad-CAM++ 히트맵

판정에 가장 치명적인 영향을 미친 픽셀 영역 시각화



베이지안 불확실성



판정 신뢰구간(90% CI) 확률 분포

모델 핑거프린터: 생성 출처의 6단계 역추적

목적: 500개 이상의 모델 레지스트리 기반 생성 출처 단일 특정

결과: 상위 5개 후보 모델의 유사도(마진)를 분석하여 점진적 압축

고급 방어 전략: 3-Class 분류와 사전적 보호

3-Class 분류 체계

기존 이분법(REAL/FAKE)을 넘어,
탐지 회피 시도의 조기 경보를 위한
ANTI-FORENSIC(반포렌식) 클래스 신설

사전 방어 (Proactive Defense)

비가시적 워터마크(Invisible Watermark) 삽입
PGD, FGSM 등 6종의 적대적 방어 기법 지원
PSNR 40dB 이상의 육안으로 구별 불가한 고품질 유지

실시간 대응 및 엣지 인프라 최적화

실시간 모니터

실시간 모니터

- 화상 회의, 라이브 방송, RTSP 스트림 대상
- 200~500ms의 초저지연 분석 및 MoltBot 즉각 알림

모델 증류 (Model Distillation)

대규모 교사 모델 지식을 소형 학생 모델로 전달 (INT8 양자화)

대규모 교사 모델
(Teacher Model)

소형 학생 모델
(Student Model, INT8)

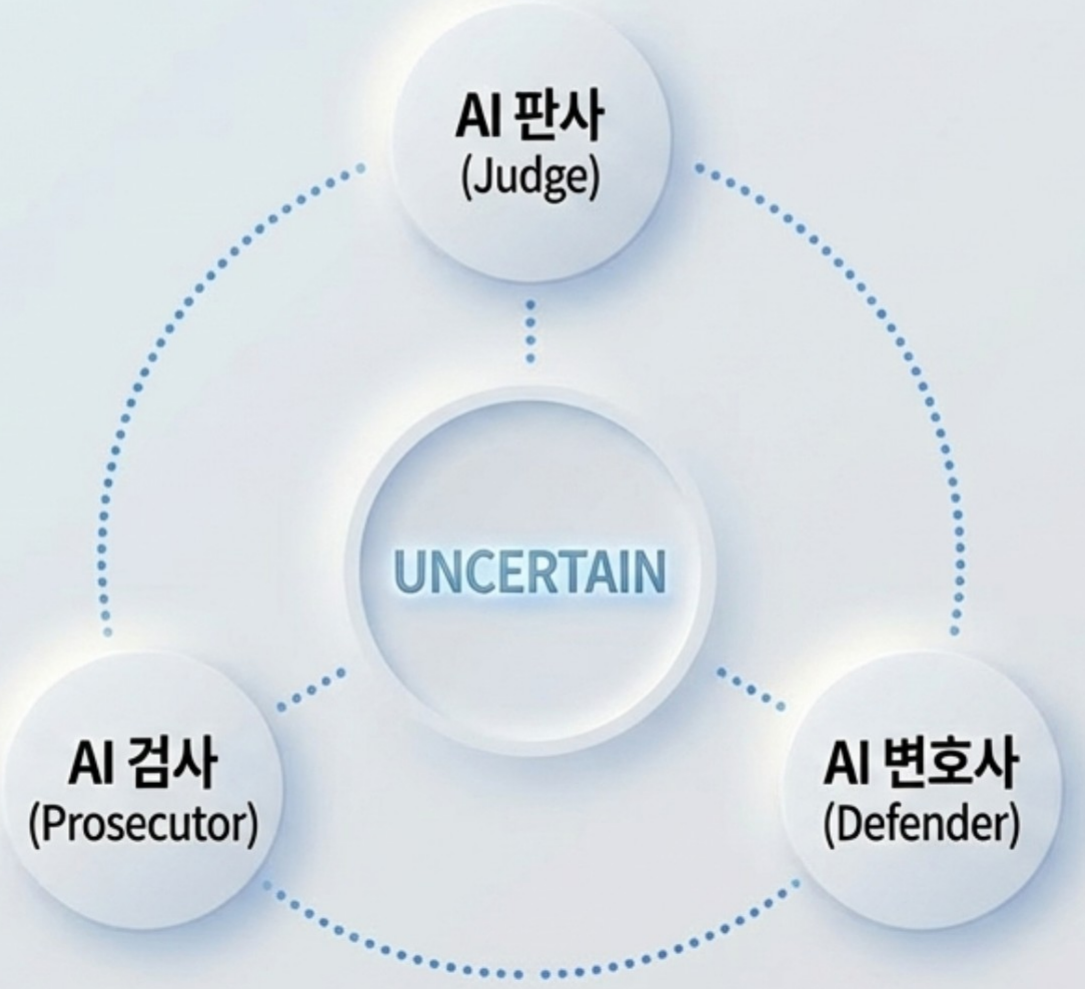
결과

- 모델 크기 1/10 축소
- 정확도 90% 유지
- 모바일, 엣지(ONNX) 지원

오케스트레이션: 멀티 에이전트 AI 시스템

다중 프레임워크

LangGraph (워크플로우),
CrewAI (역할 분담),
AutoGen (적대적 토론)



프라이버시 인프라: Ollama, vLLM

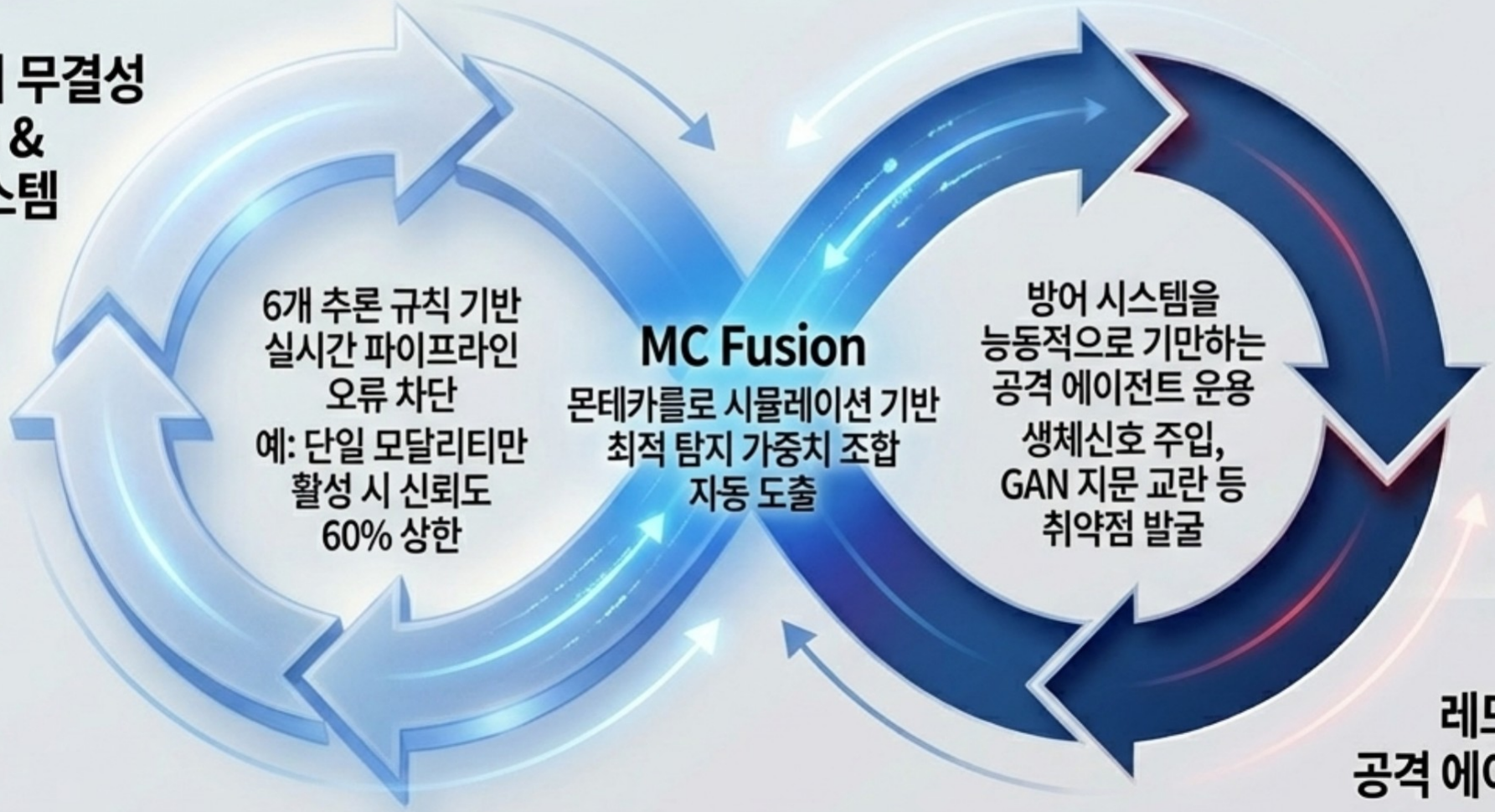
완전한 데이터 주권 유지

UNCERTAIN 케이스 해결 토론

판정이 모호할 경우
실시간 적대적 공방 진행

자기 강화적 시스템: 온톨로지 규칙과 레드팀

OWL
온톨로지 무결성
(SWRL) &
방어 시스템



ADAG
레드팀 진화 &
공격 에이전트 운용

부록 : 대한민국 법적 대응과 TruthLens의 사회적 책무

제도적 부합성 완비

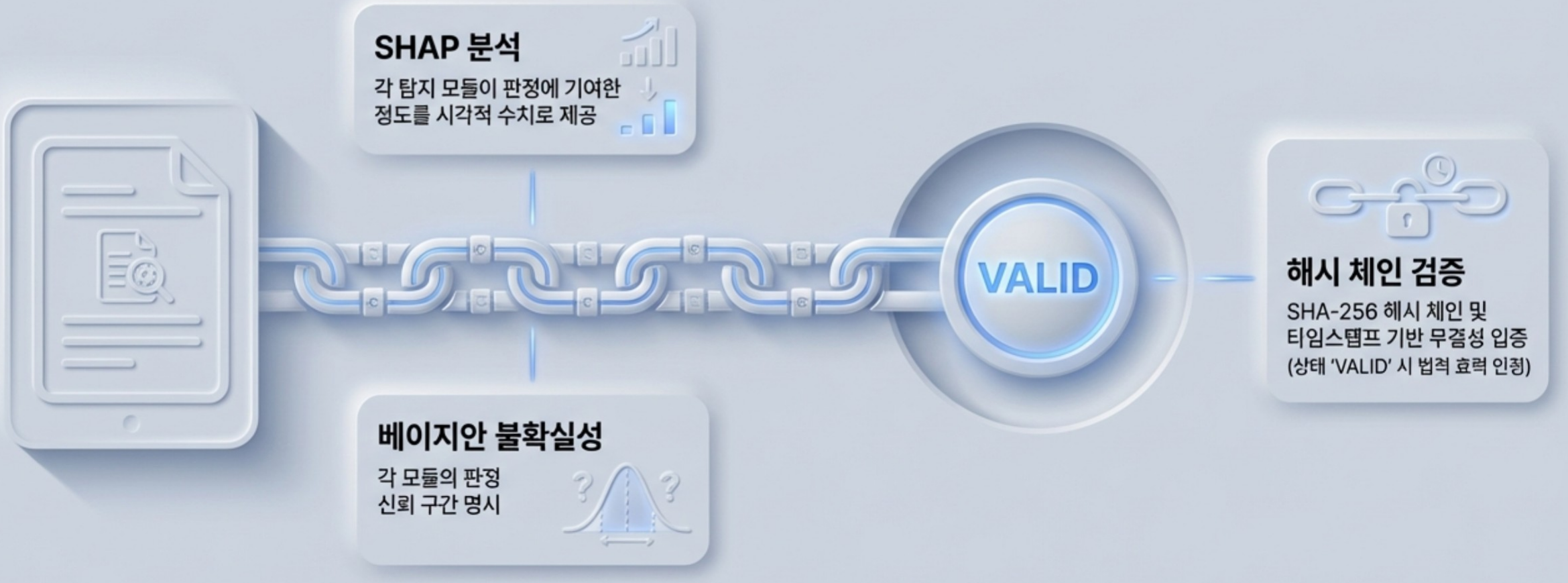
- **AI 기본법 제15조:** AI 판단 근거 설명 의무 규정 → 전 분석 건 XAI 시각화 제공
- **개인정보보호법:** 분석 종료 즉시 원본 미디어 즉시 삭제 (Zero Retention)
- **정보통신망법:** 분석 보고서(LES)를 통한 수사기관 증거 자료 지원

사회의 방패

금융 사기 방지, 민주주의 여론 왜곡 차단, 개인의 인격권 보호.
기술이 초래한 딥페이크 위협에 지능형 AI로 대응하여, 선의의 피해자가 발생하지 않도록
디지털 시대의 진실을 굳건히 수호합니다.

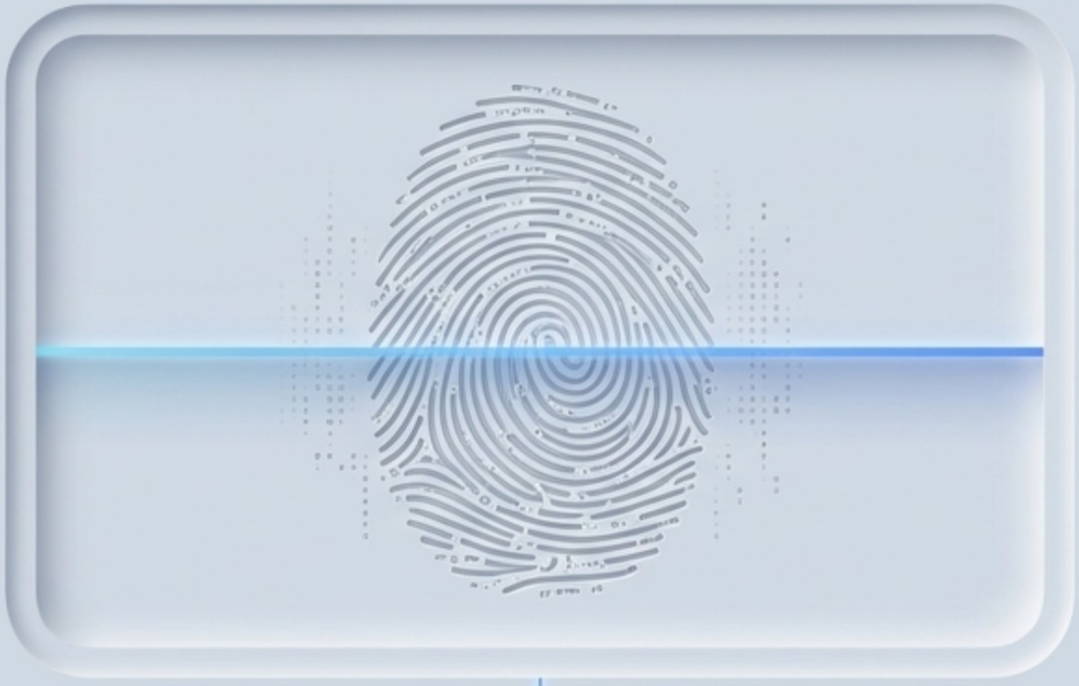
법적 증거력을 보장하는 무결성 검증 및 포렌식 보고서 자동화

분석 결과를 법적 증거력을 갖춘 국과수 및 수사기관 표준 포맷(PDF/Excel)으로 자동 생성합니다.



변형 콘텐츠의 재유포 추적과 거시적 위협 인텔리전스 매트릭스

원본의 미세 변형까지 감지하는 식별 체계와 거시적 위협 현황을 파악하는 인텔리전스 대시보드입니다.



핑거프린트 DB

변형 콘텐츠 재유포 추적

지각적 해시(Perceptual Hash) 기반 고유 패턴 추출 및 매칭

위협 인텔리전스

보안 전략 수립 지원

공격 유형(페이스스왑 등) × 대상(금융 등)의 위협 매트릭스 시각화

한국 AI 기본법 제15조 규제를 완벽히 충족하는 판단 근거 투명성 확보

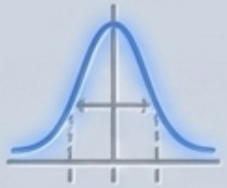
'왜 이 영상이 FAKE로 판정되었는가?' 한국 AI 기본법 제15조 'AI 판단 근거 설명 의무'를
'AI단 설명 의무'를 완벽히 충족하는 투명성 확보의 핵심 도구입니다.

SHAP 폭포 차트

각 탐지 모듈의 기여도를
앙방향 바 차트로 수치화

Grad-CAM++ 히트맵

모델이 주목한 픽셀 영역을
파란색→빨간색 열지도로 시각화



베이지안 불확실성

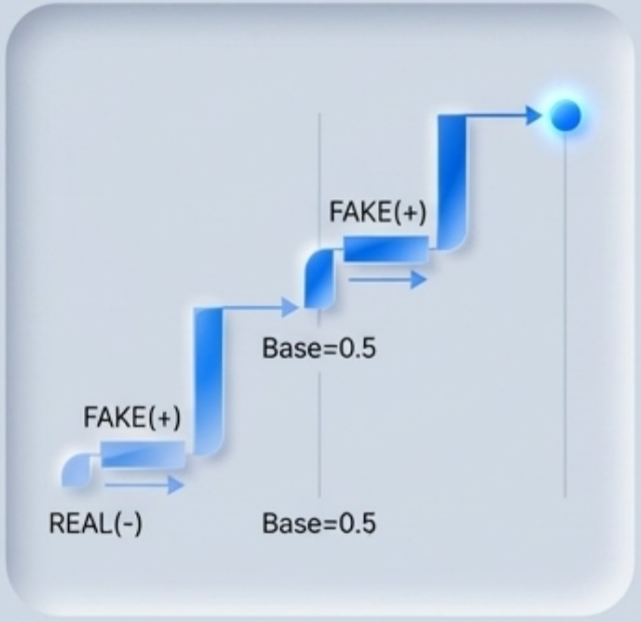
모듈별 판정 신뢰 구간(평균,
표준편차, 90% CI) 제시

모듈 기여도

탐지 모듈별 기여 비율 차트화

투명한 판독과 객관적 검증을 위한 XAI 심층 지표 해석 가이드

XAI 시각화 도구를 통해 판정 결과를 객관적으로 검증할 수 있습니다.

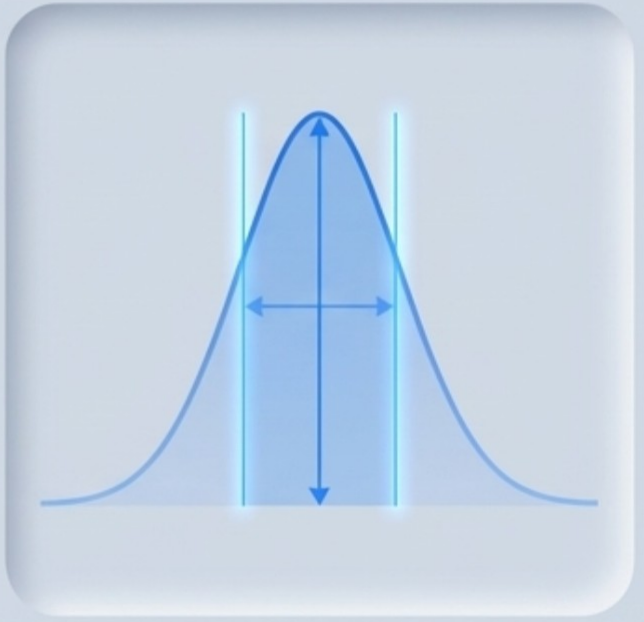


SHAP 폭포 차트 해석

기저선(Base=0.5)에서 각 모듈이 FAKE(+) 또는 REAL(-) 방향으로 이등시킨 궤적. 도달점이 최종 판정입니다.

Grad-CAM++ 히트맵 해석

빨간색 영역이 얼굴 경계부나 부자연스러운 합성 부위에 집중될수록 위조 증거가 강력함을 의미합니다.

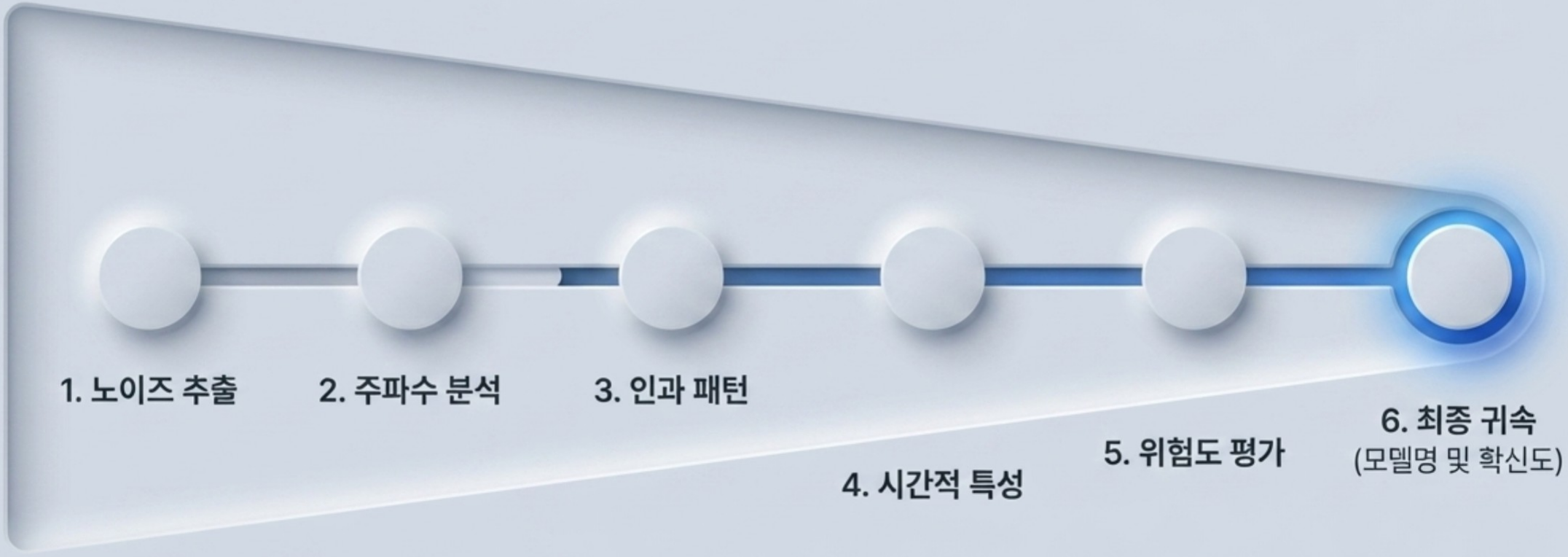


베이지안 신뢰구간 해석

신뢰구간이 좁을수록 판단의 확신도가 높으며, 0.5를 포함할 경우 해당 모듈의 증거 채택을 지양합니다.

500개 이상의 AI 모델 레지스트리를 연계한 생성 모델 역추적 파이프라인

해당 딥페이크가 어떤 생성 모델(Sora, Kling, Midjourney 등)로 만들어졌는지 역추적합니다.



전이학습(T-GD) 기반 이중 헤드 아키텍처와 법적 증거력 점수(LES) 산출

이중 헤드 아키텍처를 통해 출처를 특정하고 수사기관 제출 기준(75점 이상)을 충족하는 법적 증거력 점수(LES)를 산출합니다.

LES Score Composition

구성 요소	최대 배점	산출 기준 설명
Detection 확신도	30점	T-GD Detection Head의 진위 판별 확신도 (AUROC 95%+)
Attribution 확신도	35점	출처 모델 특정에 대한 확신도
Top-1/2 마진	20점	1순위와 2순위 추정 모델 간의 점수 격차
다중 합의	15점	기존 40개 탐지 모듈과의 판정 일치율

포렌식 탐지 회피 공격을 차단하고 원본 미디어를 보호하는 선제적 방어 체계

포렌식 조작 회피를 차단하고, 비가시적 워터마크를 통해 원본 미디어를 사전 보호합니다.

8.1 3-Class 고급 분류

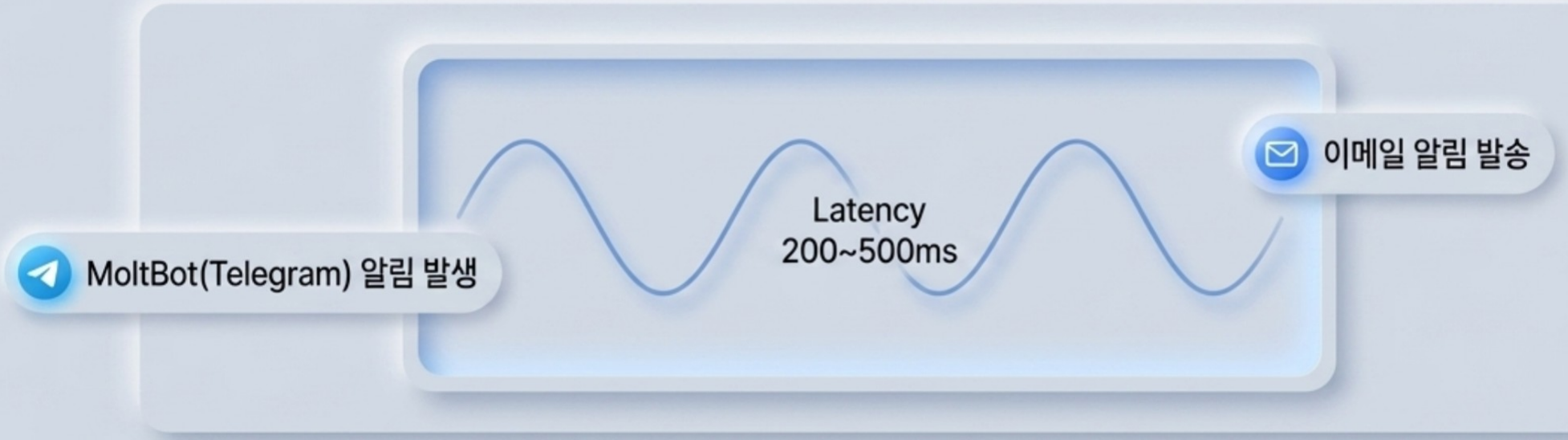
REAL, FAKE 외에
'ANTI-FORENSIC(반포렌식)'
클래스를 별도로 추가.
탐지 회피를 시도하는 고급
공격에 대한 조기 경보 발생.

8.2 사전 방어 (Proactive Defense)

비가시적 워터마크(Invisible Watermark) 삽입.
PGD, FGSM 등 6종의 적대적 공격 지원.
PSNR 40dB 이상의 고품질 설정을 통해 육안으로
워터마크 식별 불가 유지.

화상 회의 및 라이브 방송을 위한 초저지연 실시간 딥페이크 모니터링

즉각적 검증이 필수적인 라이브 환경을 위한 실시간 모니터링 시스템을 제공합니다.



입력 및 성능

웹캠 및 RTSP 스트림 지원, 초당 5~30프레임 실시간 분석 수행

초저지연 처리

Latency 200~500ms 범위 내 안정적 탐지 프로세스 유지

즉각적 알림

감지 이벤트 발생 시 MoltBot(Telegram) 및 이메일을 통한 심각도별 즉각 알림 지원

확장 감지

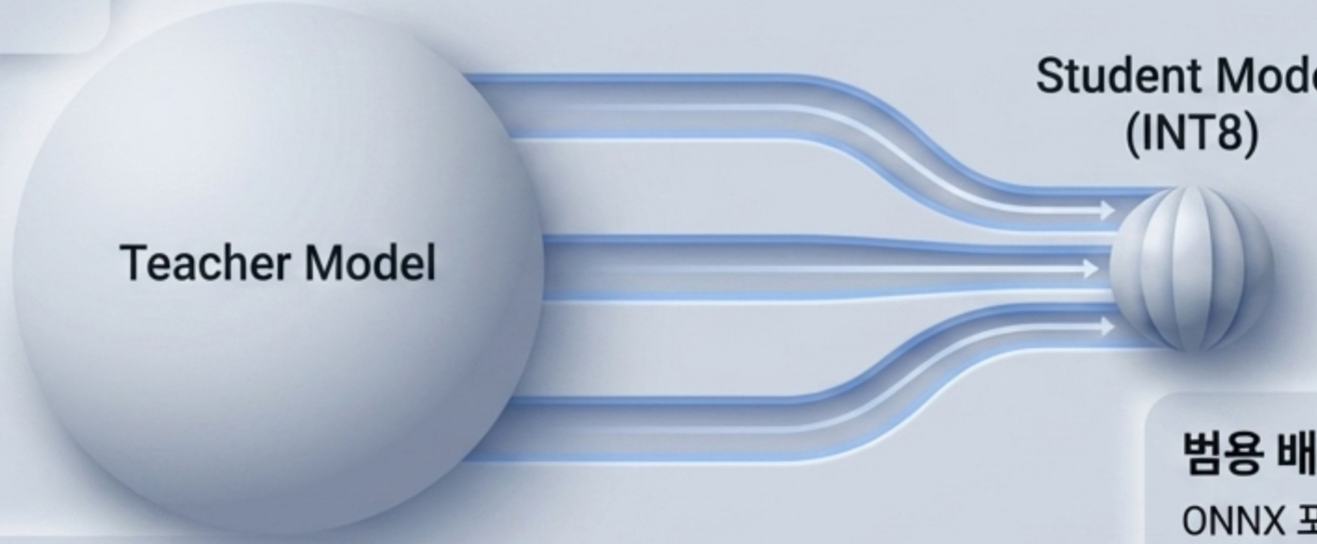
SNS 뷰티 필터 적용 여부 탐지 기능 포함

엣지 디바이스 배포를 위한 INT8 양자화 기반 초경량 모델 증류 아키텍처

서버급 대형 모델을 모바일 및 엣지 환경에 맞게 최적화하는 모델 압축 파이프라인입니다.

증류 메커니즘

Teacher 모델의 지식을 Student 모델로 효과적으로 전달.



범용 배포

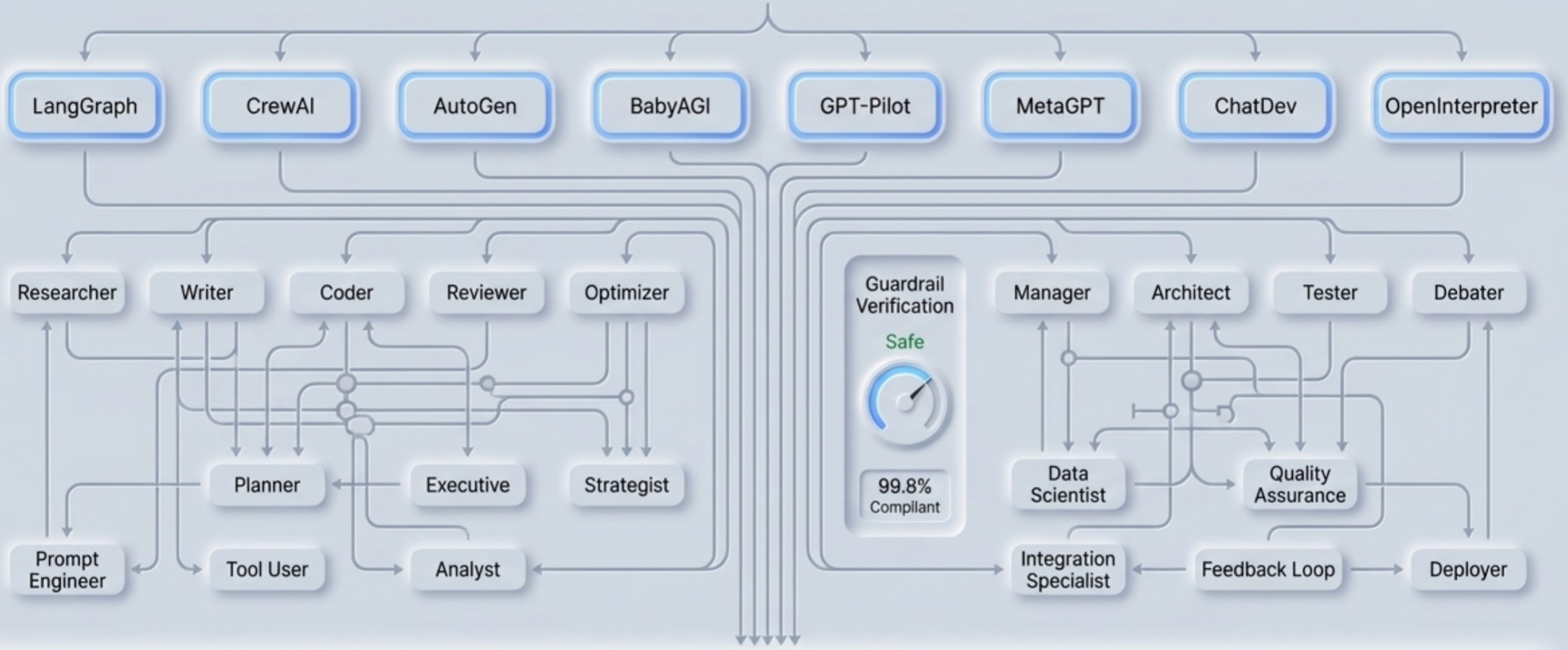
ONNX 포맷 변환을 통해 TensorRT, OpenVINO, TFLite 등 다중 추론 엔진 완벽 지원.

초경량화 기술

INT8 양자화를 결합하여 탐지 정확도 90% 이상을 유지하면서 모델 크기를 최대 1/10로 압축.

8개의 독립 프레임워크가 앙상블 합의를 도출하는 멀티에이전트 오케스트레이션

단일 모델의 한계를 극복하기 위해 8개의 멀티에이전트 시스템이 독립적으로 판정하고 앙상블 합의를 도출합니다.



프레임워크 및 에이전트 인프라

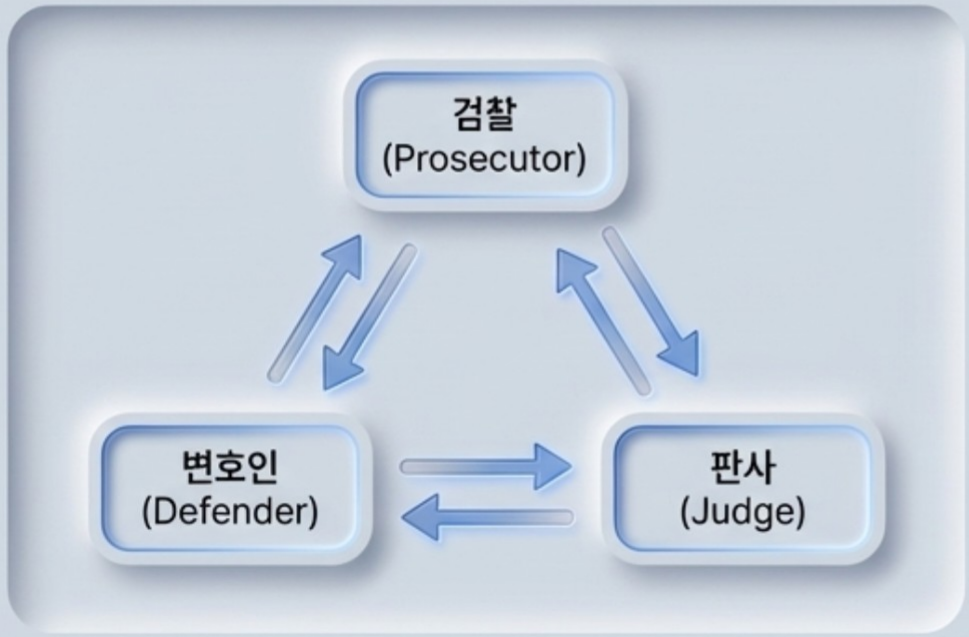
LangGraph(워크플로우), CrewAI(역할 기반 팀), AutoGen(적대적 토론) 등 8개 시스템 통합. 총 20개의 에이전트 계층 구조 모니터링.

정밀 제어

활성화 상태 제어, LLM 할당, Temperature/Context Window 등의 세밀한 파라미터 조정을 통한 속도/정확도 튜닝.

UNCERTAIN 판정을 해소하는 AI 적대적 토론과 완벽한 데이터 주권 추론 인프라

판정 유보 건에 대한 AI 간 치열한 논쟁과, 온프레미스 기반의 강력한 자체 추론 엔진 인프라입니다.



검찰, 변호인, 판사 역할 에이전트의 실시간 적대적 토론을 통한 결론 도출 (전문가 검토용 문서화).

추론 엔진: 완전한 데이터 주권을 보장하는 Ollama(로컬) 및 vLLM(서버) 환경 완벽 지원.

VLM 설정: 비전-언어 모델(기본 llama3.2-vision) 독립 설정으로 시각적 증거의 의미적 해석 강화.

시스템 취약점을 스스로 탐색하고 탐지 모델을 진화시키는 자기 강화형 적대적 훈련 (ADAG)

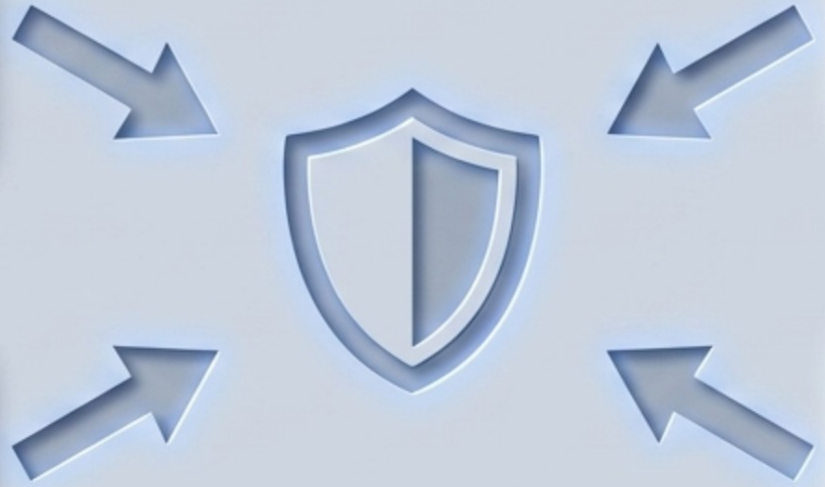
내부 취약점을 능동적으로 파악하는 Adaptive Defense-Attack Game 기반의 레드팀 인프라입니다.

BiologicalSignalInjector

미세 표정 및 눈 깜빡임 등
생체신호 주입 (현 최고 회피율)

GANFingerprintDisruptor

주파수 도메인 상의 GAN
고유 격자 노이즈 의도적 교란



TemporalManipulator

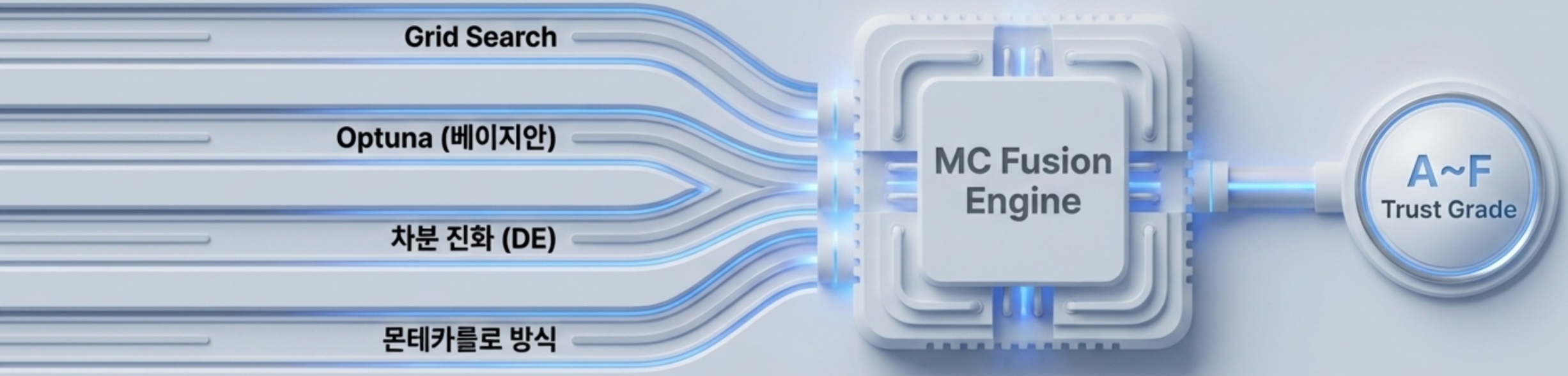
프레임 스티칭 구간의
시간적 단절선 조작 및 은폐

TextHumanizer

AI 생성 텍스트 구조의
인위적 인간화 처리

4중 병렬 시뮬레이션을 통한 탐지 모달리티 최적 가중치 조합 자동 산출

4가지 병렬 시뮬레이션 기법을 통해 딥페이크 탐지 모달리티의 최적 가중치 조합을 산출합니다.



가중치 추천

시각, 오디오, 생체, A/V 동기화 간 앙상블 가중치 자동 추출.

통합 연동

도출된 최적 가중치는 온톨로지 파이프라인의 Fusion Weights로 자동 반영되어 전체 시스템 신뢰도 등급(A~F) 상승 유도.

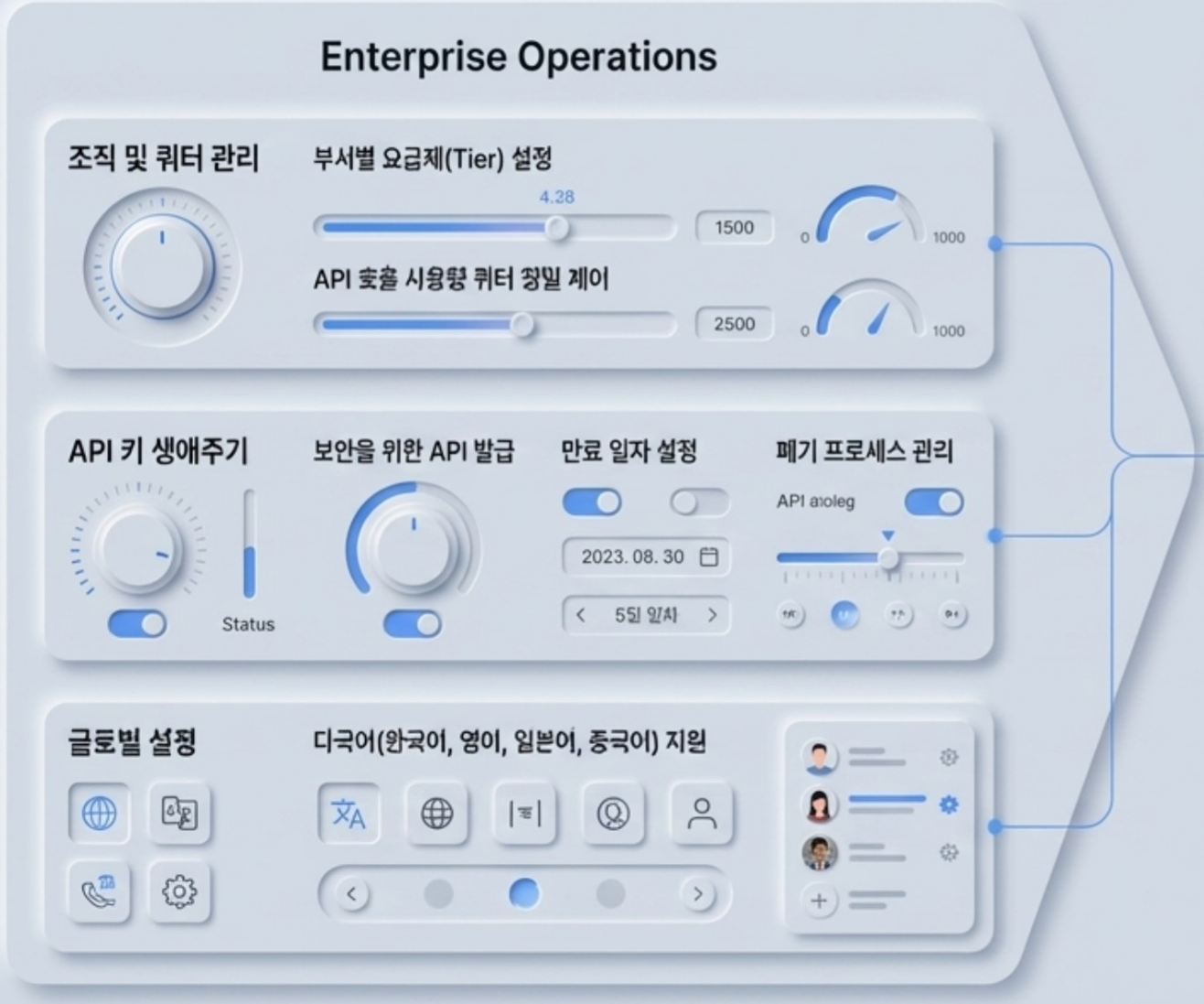
14개 노드의 무결성을 실시간으로 통제하는 OWL 온톨로지와 SWRL 추론 규칙

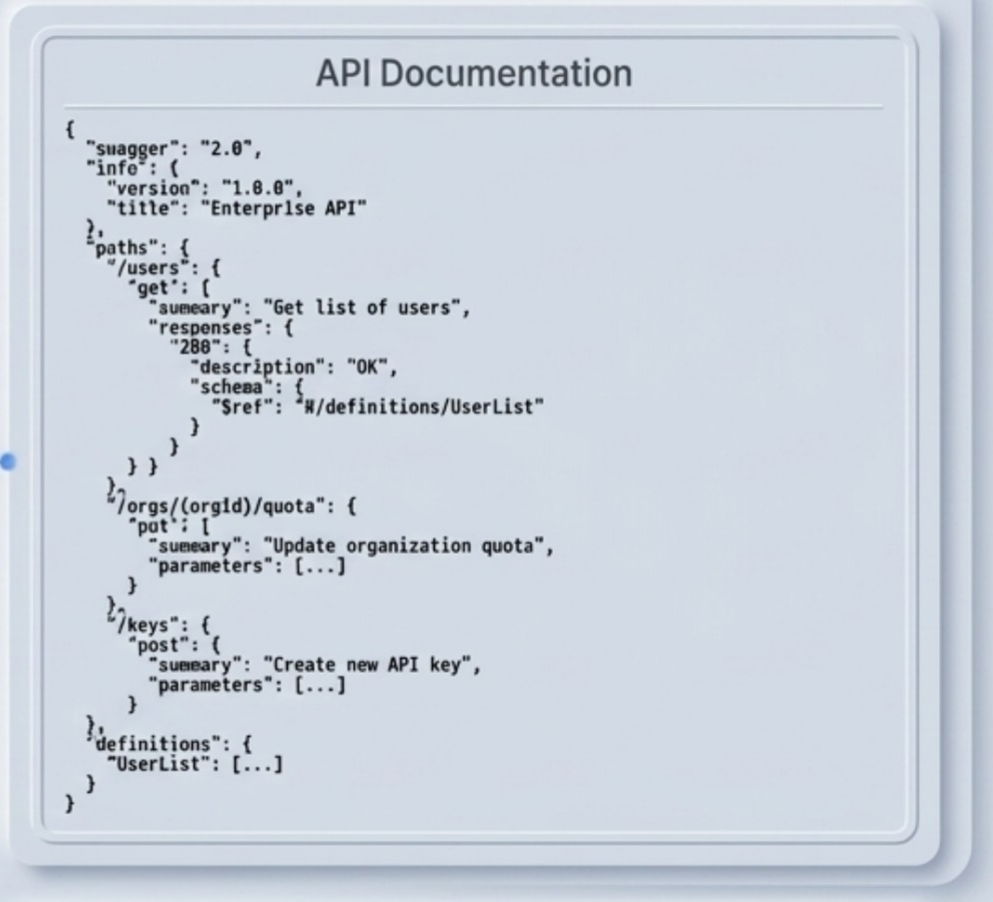
40개 모듈과 14개 노드를 OWL 형식으로 정의하고 지식 기반 아키텍처의 무결성을 실시간 제어합니다.

규칙	감지 대상	의미 및 자동 조치
SWRL-1	단일 모달리티 의존	1개 모달리티만 활성화 시 최종 신뢰도를 60%로 상한 제어
SWRL-2	데이터 순서 위반	파이프라인 노드 간 잘못된 데이터 흐름 즉각 차단
SWRL-3	VLM 환각 현상	모든 카테고리에 동일 점수를 부여하는 비전 모델 이상 감지
SWRL-4/6	의존성 및 가중치 오류	모듈 간 무한 루프 차단 및 가중치 합계 100% 무결성 확보

엔터프라이즈 환경을 위한 세밀한 조직 관리와 외부 확장을 위한 완벽한 API 인프라

관리자(ADMIN) 전용의 엔터프라이즈 운영 제어 및 외부 시스템(SI) 확장 환경입니다.





API 문서화: 외부 시스템 개발자를 위한
Swagger UI 기반의 완벽한 REST API 사양 제공.

오류를 최소화하고 신뢰성을 보장하는 과학적 판정 기준과 다각적 교차 검증 임계값

최종 확률 조건	최종 판정	해석 및 배지 색상
≥ 0.65	FAKE (위조)	조작 흔적 유의미함 (빨간색)
0.25 ~ 0.65 사이	UNCERTAIN (유보)	증거 불충분 / 모호함 (노란색)
≤ 0.25	REAL (진본)	조작 흔적 미발견 (초록색)

[신뢰도 상한 제어 정책]

1개 모달리티 교차(시각만 의존):
최대 신뢰도 60% 제한

2개 모달리티 교차:
최대 신뢰도 75% 제한

3개 이상 다각화 교차 검증 시:
100% 한도 개방

시스템 한계를 보완하고 최종 정확성을 극대화하는 4단계 휴먼 인 더 루프 (HITL) 프로세스

판단 유보(UNCERTAIN)는 시스템 오류가 아닙니다. 모호한 증거에 대처하는 4단계 권장 검증 프로세스입니다.

Step 1. 영상 품질 확인

과도한 압축 및 해상도 저하 여부 점검

Step 2. 정밀 모드 재분석

신속/표준 모드에서 정밀(Precise) 모드로 전환하여 심층 분석 수행

Step 3. 전문가 휴먼 리뷰

XAI 도구 및 AutoGen 토론 결과를 포렌식 분석관에 이관하여 해석

Step 4. 원본 파일 대조

가능한 경우 메타데이터 및 원본 미디어 확보 후 직접 비교 검증

생성형 AI의 대중화가 촉발한 전례 없는 스케일의 경제적, 정치적, 사회적 위협

생성형 AI의 대중화는 기업, 정부, 개인을 향한 전례 없는 보안 위협을 촉발했습니다.

금융 피해의 대규모화

기업 임원 화상 위조(BEC)로 수백억 원대 송금 유도 및 고도화된 음성 합성 보이스피싱 급증

정치적 의사결정 왜곡

선거 기간 중 조작 영상 유포를 통한 민주주의적 여론 조작 및 흔란 야기

미디어 신뢰도 붕괴

SNS를 통한 가짜 뉴스 확산, 유명인 사칭 사기, 그리고 개인 인격권을 침해하는 치명적 범죄 확산

대한민국의 최신 AI 규제 프레임워크와 개인정보보호 컴플라이언스 원천 준수

TruthLens는 기획 단계부터 국내외 규제 프레임워크를 원천적으로 준수하도록 설계되었습니다.

AI 기본법 제15조 준수

AI 시스템의 결정 근거 설명 의무를 XAI 심층 분석 및 무결성 감사 로그로 완벽 이행.

정보통신망법 지원

허위 영상물 유포 단속을 위해 법적 증거력(LES)을 갖춘 포렌식 분석 보고서 자동 연계.

개인정보보호법 (GDPR 포괄)

분석이 완료된 미디어 데이터의 즉각적이고 영구적인 삭제 메커니즘 적용.

기술적 탐지를 넘어, 디지털 시대의 진실을 수호하는 사회 안전망 인프라

단순한 기술 플랫폼이 아닌, AI 시대 사회 안전망의 핵심 인프라를 구축 합니다.

진실의 보호

시각·청각·생체의 다각적 교차 검증으로 디지털 허위 정보로부터 시민 권리 방어.

피해의 사전 예방

딥페이크 사기, 보이스피싱 등 악의적 공격을 조기 탐지하여 선의의 피해자 발생 원천 차단.

투명성과 진화

블랙박스를 해소하는 XAI 투명성, 그리고 ADAG 레드팀을 통해 새로운 공격에 대응하는 자기 진화형 플랫폼.

기술은 양날의 검입니다.
AI가 만들어낸 위협에는 고도화된 AI로 대응해야 합니다.
디지털 시대의 진실을 지키는 가장 견고한 방패가 되겠습니다.

11 & 12. 관리 및 시스템 섹션

플랫폼 운영 및 인프라 제어를 위한 권한별 워크스페이스

관리 섹션 (Admin)

(관리자(ADMIN) 역할 전용)

관리자 대시보드 (/admin)

사용자 목록 관리, 역할 변경, 비밀번호 초기화

조직 관리 (/admin/organizations)

조직 생성/수정, 요금제(Tier) 변경, 사용량 쿼터 설정

API 키 관리 (/admin/api-keys)

API 키 발급/폐기, 만료 일자 설정

시스템 섹션 (System)

(전역 인프라 및 연동 설정)

설정 (/settings)

언어 선택 (한국어 / 영어 / 일본어 / 중국어)
사용자 프로필 및 비밀번호 변경

API 문서 (/api-docs)

TruthLens REST API 전체 사양 제공 (Swagger UI)
외부 시스템과의 연동 개발 시 참조 지침서

13.1 판정 기준: 글로벌 임계값 (Thresholds)

딥페이크 탐지 결과의 명확한 해석을 위한 3단계 분류 기준

[0.00 ~ 0.25] 영역: REAL (진본)

최종 확률: ≤ 0.25

판정 의미: AI 생성 또는 조작의 흔적이 발견되지 않은 안전한 미디어

[0.25 ~ 0.65] 영역: UNCERTAIN (판단 유보)

최종 확률: 0.25 < 최종 확률 < 0.65

판정 의미: 판단을 내리기에 증거가 불충분하거나 모호함. (전문가 리뷰 필요)

[0.65 ~ 1.00] 영역: FAKE (위조)

최종 확률: ≥ 0.65

판정 의미: AI 생성 또는 조작의 흔적이 유의미하게 발견된 위험 미디어

13.2 신뢰도 상한 규칙 (Confidence Ceilings)

모달리티 교차 검증 수에 따른 시스템 확신도 제어 메커니즘

100%

75%

60%

[Step 3: 다중 모달리티 (시각+음성+생체 등)]
활성 모달리티 수: 3개 이상
최대 신뢰도 상한: 100%
의미: 다각적이고 입체적인 교차 검증이 이루어짐.
법적 증거로 채택 가능한 수준의 확신도 도달.

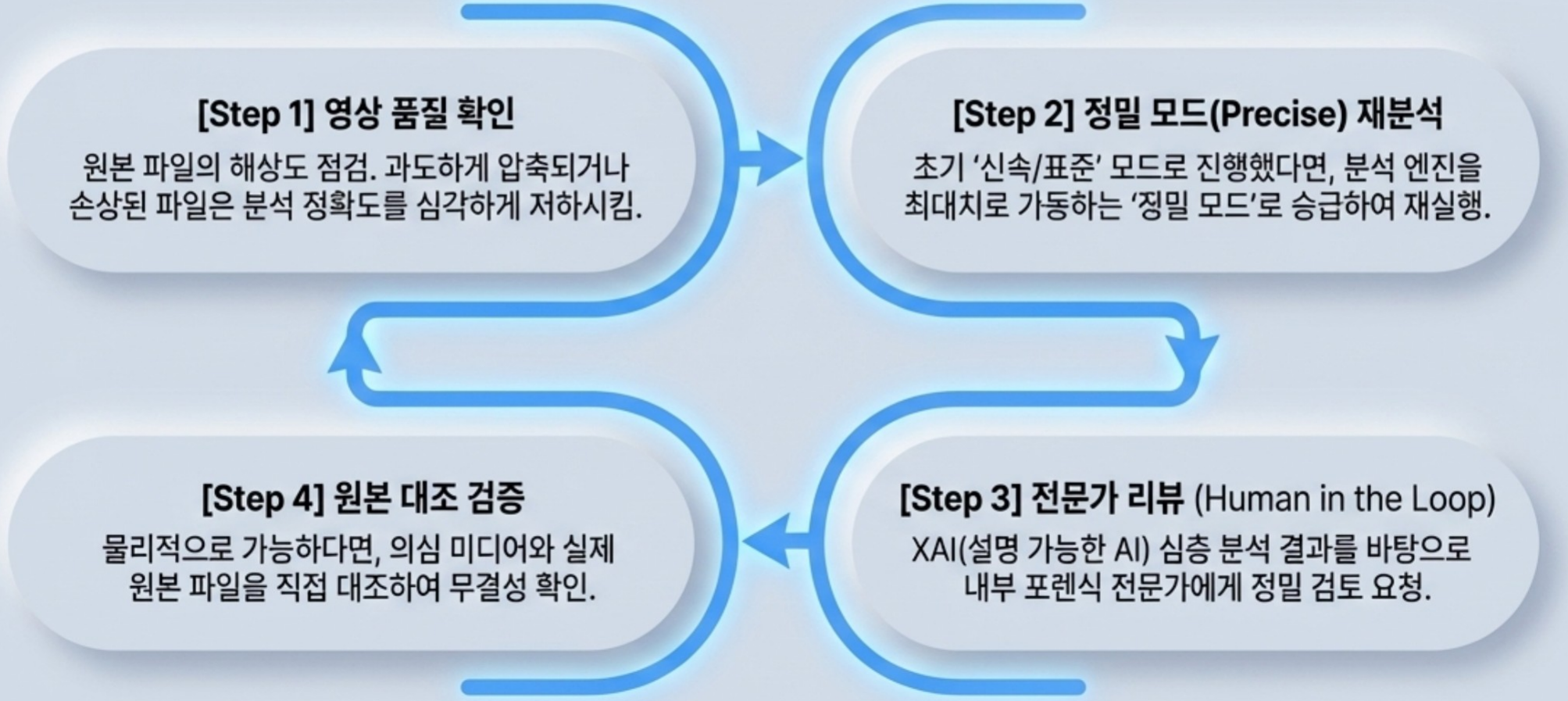
[Step 2: 이중 모달리티 (예: 시각 + 음성)]
활성 모달리티 수: 2개
최대 신뢰도 상한: 75%
의미: 두 가지 관점에서 교차 검증되었으나 완전한
확신에 도달하기엔 부족함.

[Step 1: 단일 모달리티 (예: 시각만)]
활성 모달리티 수: 1개
최대 신뢰도 상한: 60%
의미: 한 가지 관점만으로는 확신하기 어려움. 완벽한
탐지라도 시스템 오만을 방지하기 위해 60%로 제한.

13.3 UNCERTAIN 판정 시 권장 조치

모호한 분석 결과 도출 시 작동하는 대응 체계 4단계

(UNCERTAIN 판정은 분석 실패가 아닌 증거의 불충분을 의미합니다)



14.1 딥페이크 위협의 현실 (Threat Landscape)

대중화된 딥페이크 기술이 야기한 전례 없는 피해 확산

금융 분야 (Financial Fraud)

BEC (비즈니스 이메일 침해): 기업 CEO 화상 회의 영상을 위조하여 6,200만 달러(약 830억 원) 송금 유도 (2024년 홍콩 사례).

보이스피싱 진화: AI 음성 합성을 활용해 가족의 목소리를 완벽하게 모사한 지능형 사기 범죄 급증.

정치 / 사회 (Political & Social)

민주주의 훼손: 선거 기간 중 후보자의 딥페이크 영상이 유포되어 유권자의 합리적 의사결정 왜곡.

인격권 침해: 딥페이크 포르노그래피 등 일반인 및 공인의 초상을 악용한 디지털 성범죄 및 사회적 문제 대두.

방송 / 미디어 (Broadcast & Media)

가짜 뉴스 확산: 정교하게 조작된 뉴스 영상이 SNS를를 통해 급속히 퍼지며 사회적 혼란 야기.

소비자 기만: 유명인과 인플루언서를 사칭한 딥페이크 사기 광고로 인한 대규모 금전적 피해 유발.

14.2 대한민국의 법적 대응 (Legal Framework)

국가 차원의 규제 법안과 TruthLens의 완벽한 규제 준수

[AI 기본법 제15조]

법적 요구: AI 시스템 판단에 대한 '설명 의무' 규정.

TruthLens 대응: 모든 판정에 대해 Grad-CAM++ 및 SHAP 기반의 **XAI(설명 가능한 AI) 시각화**와 **해시 체인 감사 로그**를 제공하여 의무 완벽 충족.

[정보통신망법]

법적 요구: 허위 영상물(딥페이크)의 제작 및 유포 엄격히 금지.

TruthLens 대응: 탐지된 분석 보고서에 법적 증거력(LES) 점수를 부여하여 **수사기관의 공식 증거 자료**로 즉시 활용 가능하도록 지원.

[개인정보보호법]

법적 요구: 민감한 생체 및 개인정보의 무단 수집 및 활용 제한.

TruthLens 대응: **Zero-Retention 원칙**. 분석 대상 미디어를 처리 즉시 메모리에서 영구 삭제하며, 어떠한 개인정보도 시스템에 별도 저장하지 않음.

14.3 TruthLens의 사회적 책무 I

사회 안전망의 핵심 인프라로서 기능하는 3대 가치

TruthLens는 단순한 기술 도구가 아닙니다. 디지털 시대의 진실을 수호하는 방패입니다.

1. 진실 보호 (Protect Truth)

영상, 음성, 문서의 진위를 교차 검증 알고리즘으로 과학적으로 입증
고도화되는 허위 정보 및 가짜 뉴스로부터 시민사회의 인식 보호

2. 피해 예방 (Prevent Harm)

금융 사기, 지능형 보이스피싱, 여론 조작 등 악의적 목적의 딥페이크 탐지
위협 인텔리전스 및 사전 모니터링을 통해 선의의 피해자 발생 원천 차단

3. 증거 보전 (Preserve Evidence)

SHA-256 해시 체인으로 보호되는 불변의 감사 로그(Audit Log) 생성
수사 및 재판 과정에서 법적 효력(LES)을 갖춘 과학적 포렌식 증거 제공

14.3 TruthLens의 사회적 책무 II

투명성 제고와 적대적 방어 메커니즘을 통한 지속가능성

4. 투명한 판단 (Transparent Judgment)

블랙박스 해소: AI의 복잡한 연산 과정을 XAI 심층 분석 도구로 시각화

인간 중심 AI: 시스템이 독단적으로 결정하지 않으며, 구체적 근거 제시를 통해 인간 전문가의 최종 판단(Human-in-the-loop)을 완벽히 지원

5. 지속적 진화 (Continuous Evolution)

자기강화 메커니즘: ADAG 레드팀(Red Team) 시스템이 내부 취약점을 능동적으로 공격하고 보완

위협 선제 대응: 날로 교묘해지는 최신 딥페이크 생성 모델과 반포렌식 기법에 맞서 탐지 파이프라인 지속 업데이트

TruthLens: 디지털 시대의 진실을 지키는 방패

기술은 양날의 검입니다.
AI가 만들어낸 위협에는, 반드시 AI로 대응해야 합니다.
TruthLens는 디지털 시대의 진실을 지키는 방패이며,
선의의 피해자가 더 이상 발생하지 않도록 하는 것이
우리의 가장 무거운 사회적 책무입니다.

TruthLens v4.4.0

AI 기반 딥페이크 탐지 플랫폼 운영 매뉴얼

본 가이드는 TruthLens v4.4.0 기준으로 작성되었습니다.
작성: Brian Lee | AI R&D Center | A3 Security Co.,Ltd.
작성일: 2026년 4월 14일
문서 버전: Rev 4.1

문서 끝