TheNoun.

시장·사업 진입 검토 보고서

# 한국 컨택센터 Agentic Copilot

대표 가설 4개 검증 · 진입 전략

---

**담당 AI 에이전트**



**하린(Thenoun-main)**

AI Agent

총괄·검수·보고서/PPT 제작



**수아(business-research)**

AI Agent

시장·기술 리서치(조사·분석·보고서 초안)

더 나은(TheNoun)의 AI 에이전트가 조사·작성하고, 최종 판단은 대표가 합니다.

기준일 2026-09-30

발행 더 나은(TheNoun)

TheNoun.

# 한국 컨택센터 Agentic Copilot — 대표님 보고 1쪽

작성 수아(business-research) · 검수·요약 하린(Thenoun-main) · 기준일 2026-09-30 · 원문 report_v2.md(인용 147개 URL, 2차 독립 검수 통과)

## ① 결론

- **전면 AICC 신규 구축이 아니라, 기존 콜센터 시스템에 붙는「보안 통제 + 필요할 때만 AI 호출 + 상담사 승인형 업무지원」모듈로 조건부 진입.** 요약·RAG·온프렘·마스킹은 이미 국내 경쟁 제품(LG CNS·브리지텍·페르소나 등)에 있어 기능만으로는 차별화 불가.
- **가설 4개 모두 부분적으로만 맞음.**
  - ① 보안 우선 수요:「현장에 AI가 전혀 없다」는 틀림(KT 믿:음 → 케이뱅크 상담 답변추천·요약, 2023~). 그러나「프라이빗 Vertex/Azure OpenAI가 있어도 PII 우려로 못 쓰는 고객이 있다」는 기각되지 않음 → 인터뷰로 확인할 가설.
  - ② PII 가둬 두고 호출: 방향은 맞지만 가명·토큰화는 **규제 면제가 아님**(가명정보도 개인정보). 설계로 위험을 줄이는 수단.
  - ③ JEV 게이팅: JEV 자체는 공개 온프렘 모델이 아님(호스팅 API·증류 금지 약관). 대안은 로컬 소형 모델에 선택지 강제 출력 + 확률값(logprob), 한국어 인코더 분류기(KoELECTRA 등). 라우터 프레임워크만으로는 대체 불가.
  - ④ 선택적 STT·NEVA: PC 소리(고객)/마이크(상담사) 분리는 재생음이 섞여 화자 분리가 보장되지 않음. 이미 화면에 STT가 뜨는 곳의 NEVA 화면 읽기는 가장 싼 경로지만 실시간 완전성 검증 필요.
- **시장 규모는 확정치 없음.** 상담 인력 12만~40만 명(2022 총람 집계~협회 추정) × 월 3만~10만원 가정 → 연 432~4,800억원은 **민감도 범위일 뿐**. 승인·예산이 있는 계정 기준 SAM은 **36~144억원** 가정. 가격 기준점: 해외 Zendesk Copilot 상담사당 월 50달러.

## ② 대표님이 결정할 것

1. **90일 제한형 PoC 착수 여부 (추천: 착수).** 대상 = 기존 STT·업무 접근권이 있는 민간 고객사 + BPO 2곳. 범위 = 상담사 승인형 검색·요약·스크립트 지원까지(자동 거래·신규 TTS·STT 교체 제외). 예산 가정 24인주 · 약 6,300만~1.2억원(견적 미확인).
2. **현대차 현장 PoC — 대표님 판단으로 PoC 현장 확보 가능.** 가설 ①④·NEVA 연동을 가장 직접 검증할 수 있는 곳.
3. **자체 모듈 판매 vs 기존 벤더 통합 파트너** — PoC 결과(비용·품질·구매 의향) 보고 결정. 금융은 승인 확보된 파트너 경로, 공공은 초기 제외.

## ③ 알아서 할 것 (하린·수아)

- PoC 설계서: 한국어 개인정보 탐지·게이트 성능 목표와 중단 조건(중대 PII 외부 노출·타 상담 복원·권리 미확인 시 즉시 중단).
- 로컬 판단 모델·STT 후보의 라이선스 재확인(A.X 4.0 표준은 Apache 아님, Kanana MoE 온프렘 재판매 별도 허가, GLiNER-ko 비상업 등).
- AI기본법(최초 시행 2026-01-22 · 현행 법률 2026-07-21 · 시행령 2026-08-20)·통신비밀보호법(통화 수집 고지·동의)·상담사 개인정보 검토 목록 정리.
- 미확인 채우기: 최신 전국 좌석 수, 국내 단가 견적, 심의 일정, NEVA 버전·라이선스.

TheNoun.

# 대표님 가설(원문 요지)

1. **온프레미스·보안 우선:** 금융·공공·대기업 콜센터는 Private Network 로 구성한 Vertex AI 나 Azure OpenAI Service 로 AI 를 공급받을 수는 있어도, 개인정보 유출이 우려돼 상담 현장(콜센터)에는 AI 를 못 내보내고 있다.
2. **개인정보를 가둬 두고 호출:** 개인정보는 사내에 가둬 두고, AI 에는 가명·토큰만 보내 호출한 뒤 응답을 사내에서 되돌려 끼우는 방법론(대표가 연구해 온 방식)이 해법이다.
3. **JEV(오픈소스 필요)로 빠른 판단 → 필요할 때만 비싼 호출:** 고객과의 대화를 실시간으로 보면서 JEV 같은 빠른 판단 모델이 판단하고, 항상 AI 를 부르는 게 아니라 필요할 때만 비싼 LLM 을 호출한다. 온프렘에 올리려면 오픈소스여야 한다(JEV의 오픈소스 대체제로 LAYA 가 있다).
4. **STT·TTS 를 무조건 붙이지 않는다:** PC 로 들어오는 소리 = 고객, 마이크 = 상담사로 나눠 STT 할 수 있고, 이미 STT 결과가 화면에 뜨는 곳은 RPA(알피에이(주) NICE NEVA)로 화면 값을 실시간으로 받아 쓸 수 있다.

TheNoun.

# 전체 보고서

## Executive Summary — 대표님 의사결정용 1쪽

### 결론 3줄

1. **[의견] 전면 AICC 신규 구축보다 기존 시스템에 붙는 '보안 통제·선택 호출·승인형 업무지원' 모듈의 조건부 진입 권고.** 요약·RAG·온프렘·마스킹은 이미 경쟁상품에 존재하므로 그 기능만으로 시장 공백 주장 불가.[57][68][85]
2. **[사실/판정] 네 가설 모두 일부 유효하나 그대로 사업 전제 채택 불가.** KT·케이뱅크 믿:음 사례는 현장 생성형 상담 지원의 전면 미도입을 반증하지만, 프라이빗 Vertex AI/Azure OpenAI가 있어도 특정 고객이 PII 우려로 도입하지 못한다는 가설까지 기각하지 못함.[87][4]

   Jev 자체는 공개 온프렘 모델로 확인되지 않으며, PC 두 소스·NEVA 화면은 역할·실시간 완전성 검증 필요.[45][28][48]
3. **[의견] 90일 PoC는 STT API·로컬 판단·사내 PII 경계·읽기/권고부터 시작.** 중대 PII 외부노출·타 상담 복원·권리 미확인 발생 시 즉시 중단, 비용·품질 개선과 구매확약 없으면 확장 보류. 구체 문턱은 아래 제안이며 실측 아님.

### 대표님이 결정할 것

- **추천:** 기존 STT와 업무접근권 가진 민간 고객사+BPO 2개 후보를 대상으로 제한형 PoC 준비. 금융은 용도별 승인 확보된 파트너 경로, 공공은 초기 SOM에서 보수적으로 제외.
- 제품범위: 상담사 승인형 검색·요약·스크립트 지원까지; 거래/고객정보 자동 쓰기·신규 TTS·STT 전면교체는 1차 범위 제외.
- 검증 후 결정: 자체 PII/게이팅 모듈 판매 vs 기존 벤더 통합 파트너; 월 3/6/10만원 등록 사용자 좌석 가격 가설과 구축비 별도 견적 실험. **확정 시장가격 아님.**

### 핵심 숫자 3개

| 숫자 | 의미·원출처·기준 | 의사결정 주의 |
|---|---|---|
| **80,027명** | 2021년 전국사업체조사 '콜센터 및 텔레마케팅서비스업'; KOSIS 답변 2024-03-20.[1] | 전체 상담사·2026 좌석 수 아님. 2022년 수치로 인용 금지 |
| **123,881명** | 2022년 총람 사용업계 집계; 총람 작성기관 프라임경제2023-08-30.[3] | 전국 전수 확정 아님; 사람이 곧 동시좌석 아님 |

| 숫자 | 의미·원출처·기준 | 의사결정 주의 |
|---|---|---|
| **432~4,800억원/년** | 인력 12만~40만명을 등록 좌석으로 가정 × 월 3만~10만원 ×12. 상·하한 **민감도 가정**, wiki/v2-calculations | 조사로 산출한 시장 추정 구간·신뢰 구간 아님. 구축·STT·모델·도달률 제외; TAM 확정치로 사용 금지 |

### 다음 검증이 바꿀 결론

- 기존 벤더가 동일 통제 경계·연동·가격을 제공하고 고객 교체수요가 없으면 독립 제품 진입 보류.
- 한국어 게이트 중요개입 recall·PII 탐지·유지보수 원가가 목표 미달이면 범위를 줄이거나 로컬/수동경로 채택.
- 최신 전국 좌석·Copilot 매출·구매의향은 **미확인**. 시장규모를 단일 수치로 확정하지 않음.

## 0. 근거 사용·범위

- **[사실]:** 확보한 법문·공식 제품문서·발표·연구 소개의 내용. 벤더가 '보장한다'는 문구도 공급자 주장이지 독립 실험결과가 아님.
- **[추정]:** 명시한 계산식·모집단·가격/도입 가정. **[의견]:** 진입·설계·가격 실험·PoC 통과기준.
- 공공/금융 법률 적용: 고객 법무·감독기관·보안 검토 조건부. 일반 법문·공개 발표로 개별 서비스 적법성 인증 불가.
- 신규 고객접촉·모델 실행·현장 PoC·계약·외부 발송은 수행하지 않음. 본 산출물은 공개자료 조사 및 실험 설계.
- 과거 고객사례의 2026 지속운영과 최신 모델은 자동 승계하지 않음. 자료 발행일 미표기는 확인일 2026-09-30만 표시.

### 0.1 약어·측정 단위

- **PII:** 개인 식별 정보. 정형 ID뿐 아니라 이름·주소·음성·대화 맥락의 간접 식별 가능성 포함.
- **AICC / CCaaS:** AI 컨택센터 / 클라우드 구독형 컨택센터. Copilot 단품 시장과 구별.
- **STT / TTS / RAG:** 음성→텍스트 / 텍스트→음성 / 검색한 근거를 붙여 답변을 생성하는 방식.
- **sLLM / NER:** 소형 언어 모델 / 이름·주소 등 개체명 탐지. 작은 모델이라는 표현만으로 한국어 성능·원가 보증 불가.
- **G1 / G2 / G3:** 개입 필요 판단 / 생성 모델 선택 / 출력·행동 허용 검사. 보고서의 기능 분해이며 공인 등급 아님.
- **named-seat / concurrent-seat:** 등록 사용자 좌석 / 동시 접속 좌석. 사람·교대·물리 좌석과 구별.

- **recall / abstain:** 반드시 개입해야 할 사건을 잡은 비율(재현율) / 판단 보류 후 사람·규칙·상위 모델로 넘기는 처리.
- **AHT / ACW / QA:** 평균 통화 처리 시간 / 통화 후 정리 시간 / 상담 품질 평가.
- **RFI / RACI / SLA:** 공급자 정보 요청 / 책임·승인·협의·공유 역할 표 / 계약상 서비스 수준 약정.
- **TCO / ARR / TTFT:** 전체 소유·운영 비용 / 연환산 반복 매출 / 첫 토큰까지 시간. ARR은 첫해 인식 매출과 다름.
- **p50 / p95 / p99 / CI:** 지연 분포의 50/95/99백분위 / 통계적 신뢰 구간. 각 구간의 p95 합을 전체 p95로 계산하지 않음.
- **fail-closed / SBOM:** 불확실하면 노출·실행을 차단하는 정책 / 배포 소프트웨어·모델의 구성과 권리 목록.

# 1. 시장 정의·규모 — TAM/SAM/SOM

## 1.1 본 보고서의 시장 정의 [분석 정의]

- Copilot: 상담사의 실시간 지식검색·요약·응대권고·승인형 업무지원. Agentic은 제한된 도구/API 호출과 업무단계 조합을 포함하되 자동 거래권한과 동일시하지 않음.
- **제외:** BPO 인건비/매출 전체, CTI/IPCC 전체, CCaaS 전체, STT/TTS 전체, 고객 자동응대 봇 전체. 넓은 AICC 매출과 좁은 Copilot 매출의 합산 금지.
- 좌석기준: 등록상담사(named)·동시좌석(concurrent)·물리좌석·실사용자를 분리. 아래 시장계산은 named-seat 가정.

## 1.2 현재 확인 가능한 모집단

| 관측값 | 기준연도·출처 | 확정 가능한 범위 / 미확인 |
|---|---|---|
| 1,750개 사업체·80,027명 | 전국사업체조사 2021 / KOSIS2024-03-20.[1] | 특정 산업분류. 직영 상담인력·좌석 총계 대체 불가 |
| 123,881명 | 컨택센터산업총람 2022 집계 / 프라임경제2023-08-30.[3] | 사용업계 표본성 집계, 전수성 미확인 |
| 약 40만명 | 같은 기사에 소개된 협회 관계자 추정.[3] | 전수조사·공식 좌석 통계 아님. 기준일 현재값으로 채택 안 함 |
| 통신 16.7%·은행카드 18.8%·보험 12.2% | 총람 2022의 사용업계 분포.[3] | 합 47.7%, 표본 약 59,091명(계산). 국내 Copilot 매출점유율 아님 |
| 직접운영 44.2%·위탁기업 중 500명 이상 82% | 민주노동연구원 워킹페이퍼 2024-03 소개, 발행 2024-02-28.[2] | 고객사·BPO 구매 구조의 단서. 조사모집단 원문미확인으로 전국 일반화 금지 |

출처: KOSIS 원답변, 총람 작성기관 기사, 민주노동연구원 연구소개. 최신 2026 종사자·좌석·업종전수 분포는 미확인.

## 1.3 계산식·시나리오 [추정]

TAM_reference = N_named_assumed × P_month_assumed × 12

SAM = Σ(연동권한·예산·보안승인 충족 계정의 실제 라이선스수 × 검증된 월가격 ×12)

SOM_ARR = Σ(계약좌석 × 계약월가격 ×12)

실제 첫해 인식매출은 고객별 서비스 개시월·부분월·할인·사용량·회계정책 반영, ARR와 구분.

| 단계 | 모집단/가격 가정 | 계산결과 | 채택 수준 |
|---|---|---|---|
| TAM 참고 저/중/고 | 123,881명→named-seat 1개씩 **가정**, 월 3/6/10만원 | 연 445.97/891.94/1,486.57억원 | 전체시장 확정치 아님. 좁은 유료Copilot 도달률 미반영 |
| SAM 저/중/고 | 자격 갖춘 5천/1만/2만 named-seat **가정**, 월 6만원 | 연 36/72/144억원 | 고객계정 bottom-up 미실사, 시장조사 결과 아님 |
| SOM 저/중/고 | 최초 12개월 계약 200/600/1천 seat **가정**, 월 6만원 | 연환산 ARR1.44/4.32/7.2억원 | 판매실적·확정계약·첫해 인식매출 아님 |

출처: N의 구년도 참고는 프라임경제2023-08-30; 가격/도달/계약 가정은 수아 제안, wiki/v2-calculations의 실행 산술 결과. 구축·STT·인프라·모델·SI 별도 제외.[3]

## 1.3a 모집단 불확실성 민감도 [가정]

| 가정 인력→등록 좌석 수 | 월 3만원 | 월 6만원 | 월 10만원 |
|---|---|---|---|
| 12만 | 연 432억원 | 연 864억원 | 연 1,440억원 |
| 40만 | 연 1,440억원 | 연 2,880억원 | 연 4,800억원 |

산식: 가정 인력 × 사용자당 라이선스 1개 × 가정 월 SW 가격 ×12. **12만~40만은 신규 조사 결과가 아닌 민감도 입력 범위**. 구년도 123,881명 집계와 협회 약 40만명 추정의 불확실성을 드러내기 위한 연구자 설정. 위 범위 전체가 구매 가능한 Copilot 시장이라는 뜻이 아님. 실제 SAM은 승인·API·예산을 확보한 고객별 계정으로 계산. 원 참고 집계: 프라임경제·2023-08-30.[3]

## 1.4 성장률·자료공백

- **확정할 수 없는 것:** 한국 '상담사 Agentic Copilot'의 최신매출·보급좌석·연성장률.
- AMR Korea Call Center AI·GVR Korea CCaaS 시장자료는 검색문 단서를 찾았으나 본문 403/방법론 미확인이고 시장범위가 다름. **공개 검색문 숫자를 본 보고서 성장률·TAM으로 채택하지 않음.**

- 한국노동연구원 2025-11 연구의 금융콜센터 장은 검색문만 확인·PDF403. 비용/생산성 결론을 본문검증 근거로 인용하지 않음.
- 다음 확보: KOSIS 최신 전국사업체조사 75991 표, 고객/BPO 실제 라이선스·교대·중복집계, 최신 총람 방법론·기관별 운영좌석, 발주/낙찰 범위별 모듈 금액.

## 2. 수요·구매 동인

| 동인 | 확인 사실 | 사업 해석 [추정/의견] | 검증 |
|---|---|---|---|
| 검색·후처리 부담 | KT Assist는 실시간 대화록·지식추천·요약, FCC는sLLM/RAG 지원 공개.[59][57] | 시간절감만 아닌 오상담·탐색·교육 부담을 구매가치로 제시 | 동일업무 ACW/AHT/QA·편집률, 고객평가/상담사부담 병행 |
| 시스템 교체 부담 | 넥서스 기존IPCC/구축형과 Verint 기존CX 통합 설명.[70][79] | 교체보다 기존STT/CRM/KMS를 유지하는 모듈 선호 가능 | API·권한·조달책임자 인터뷰; 실제통합 견적 |
| 보안·감사 | 금융 SaaS 일반예외는 데이터/통제 조건부; SONA는 온프렘·마스킹·로그 공개.[8][85] | '아무도 못함'이 아닌 승인·감사 증빙 비용 절감 기회 | 심의 리드타임·반려사유·통제증빙 견적 |
| 고객사/BPO 분리 | 직접운영/위탁 구조 연구소개 존재.[2] | 비용부담자·효익수취자가 달라 판매 병목 가능 | 고객사 예산·BPO운영·SI보안 3자 RACI와 구매확약 |
| 규제변화 | 금융 생성형AI 지정/출시와 SaaS 예외, 공공 공동기반 방향 발표.[7][8][14] | 진입경로가 열리면서 신규단독시장뿐 아니라 기존SI 경쟁 확대 | 지정조건·기관공고·공동기반 호환성 확인 |

출처: 위 행별 원문. 동인별 시장 비중·지불의향·실제 절감은 미측정. 통신/금융 비중을 신생제품의 구매가능성으로 바로 환산하지 않음.

## 3. 경쟁 지형 — 국내 우선, 해외 비교

**읽는 법:** 아래 [사실] 표는 제품/도입 설명 확인. V=벤더발표, C=고객사발표, M=기업발표 전달 매체. 모두 독립 성능시험과 구별. 미공개=읽은 원문에 없음; 미확인=접근/범위/시점 문제.

### 3.1 국내 제품·역할·배포·모델

| 벤더 | 확인 제품/서비스 | 상담사 지원(Copilot) | 고객 자동응대 | 배포형태: 확인 범위 | LLM 공급원: 제품별 확인 범위 | 근거 |
|---|---|---|---|---|---|---|
| 삼성SDS | AICC (2021 발표) | 실시간 답변·제품정보·가이드 추천, 품질분석 | 챗봇/음성봇·복잡 상담 이관 | 읽은 발표에 미공개 | NLU/STT/TA 공개, 생성형 LLM 공급원 미공개. 일반 업무 Copilot을 AICC 모델로 대체 추정하지 않음 | [56] |

| 벤더 | 확인 제품/서비스 | 상담사 지원(Copilot) | 고객 자동응대 | 배포형태: 확인 범위 | LLM 공급원: 제품별 확인 범위 | 근거 |
|---|---|---|---|---|---|---|
| LG CNS | **FCC RT-Advisor**, FCC Agent Desktop; DAP Gen AI Talk; Singlex 챗봇 | RT-Advisor: 지식추천·요약·분류, Desktop 임베디드 | DAP/챗봇/음성봇 등 별도 기능 | CCaaS SaaS; 고객 전용형 private/public cloud. 모든 구성의 완전 온프레미스 제공은 이 문서만으로 미확인 | RT-Advisor: 내장 상담 전용 **sLLM**, 다른 생성형 모델 접목 가능. Singlex는 **Google Dialogflow** PaaS(LLM 이름 아님). RT-Advisor의 기본 모델 이름/버전 미공개 | [57][58](시점 미상) |
| KT | AICC **상담Assist** | 실시간 STT 대화록·AI KMS 답변추천·유형/요약·관리자 모니터링 | 보이스봇·챗봇/농협 핵심업무 연동 별도 | Assist는 솔루션 설명; 공공 사이트의 G보이스봇/G챗봇 SaaS 표시를 Assist 전체에 적용하지 않음. 농협 맞춤 구축 확인 | KT 그룹사 AICC의 **믿:음** 적용은 2025 보도(M). 상담Assist 자체는 KT 기술문서상 LLM+RAG(1·2차 파일럿 AWS Bedrock·MS Azure), 모델명·버전 미공개(§5b.1). 믿:음2.0 확대는 당시 계획이며 모든 B2B 제품의 현행 기본 모델은 미확인 | [59][87][82][192] |
| SKT | **AI CCaaS**, 자사 AICC | 실시간 지식추천, 자동 요약/분류, QA/코칭 | 챗봇·보이스봇, 24시간 운영 | 클라우드 구독형 올인원 CCaaS | 읽은 제품/운영 기사에 모델 공급원 미공개. A.X 보유 사실로 해당 제품의 기본 모델을 추정하지 않음 | [61][62] |
| 네이버클라우드 | **CLOVA AiCall** | 이 자료에서 상담사 실시간 Copilot 제품/운영 **미확인** | 현대커머셜 완전판매 해피콜 운영(C) | 고객 발표로 서비스 도입 확인. Classic/VPC/온프레미스 상세 문서는 접근 실패하여 본문 미검증 | 고객 발표는 CLOVA AiCall 명시. 당시 또는 현재 AiCall이 **HyperCLOVA X를 추론 LLM으로 사용**한다는 결론은 미확인 | [81] |
| 카카오엔터프라이즈 | 카카오 i 커넥트 센터 **SaaS형** (2023) | 사람/AI 협업·콜백·대시보드 확인, 생성형 상담사 Copilot 미확인 | 전화 음성봇의 주문·환불·배송 등 안내 | 카카오 i 클라우드 기반 SaaS, 월구독 | 2023 원문에 LLM 공급원 미공개. 현재 센터플로우 링크의 판매/주체/가격 본문 확인 실패 | [60] |
| 브리지텍 | **AoE**, Athena AICC / ForCus AI 연계 | STT→sLLM 대화요약; 상담지원 구축사례 목록 | Virtual Agent: 조회·처리, RAG/Tool/MCP | 자체 sLLM **온프레미스 설치 가능**. 나머지 제품/외부 호출의 위치는 계약별 확인 | 자체 sLLM + 외부 **GPT·Claude** 연동을 명시. sLLM 구체 모델/라이선스 미공개 | [68][69] |
| 넥서스커뮤니티 | AICC 패키지·CS플랫폼, ROVA 사례 표기 | 실시간 상담지원·STT/TA·AI 지식관리/분석 | AI 음성 상담봇·자동주문 | IPCC **클라우드/구축형** 명시 | STT/TA 기반 설명. 생성형 LLM·공급원 및 agentic Tool 실행은 미공개/미확인 | [70] |

| 벤더 | 확인 제품/서비스 | 상담사 지원(Copilot) | 고객 자동응대 | 배포형태: 확인 범위 | LLM 공급원: 제품별 확인 범위 | 근거 |
|---|---|---|---|---|---|---|
| 스켈터랩스 | **BELLA QNA** (업무규정 RAG 예시), AICC/STT 고객사례 | 사내 지식 챗봇의 업무보조 확인. 실시간 음성 Copilot과 동일한 실적 아님 | 퍼시스/요기요 AICC 자동응대 고객사례 | API·Slack 연계 확인. BELLA/AICC별 SaaS·private·온프레미스 조건 미확인 | BELLA QNA 예시는 **ChatGPT 연동**. 홈페이지 일반 맞춤 LLM/Agent 사업을 해당 사례의 모델·배포로 일반화하지 않음 | [63][64] |
| 마음AI (원문 마인즈랩) | **FAST AICC**, VOC, TMQA (기존 페이지) | 상담지원·VOC·QA/STT/패턴분류 | 음성봇·시나리오형 챗봇 | 서비스 소개/문의형, 구체 AICC 배포 조건 미공개 | 이 원문에서 생성형 LLM 공급원 미공개. 2022 푸터의 레거시 사례를 최신 M3 등으로 치환하지 않음 | [65] |
| 페르소나AI | **SONA ASSISTANT**, GEN AICC/HICC, KGPT(별도 업무지식 제품) | STT 기반 RAG/LLM 추천·분석·평가·리스크 탐지·AI 도우미 | GEN AICC/HICC AI+사람 하이브리드 | ASSISTANT **온프레미스 보안 인프라**; GEN AICC 사용량 과금. 상품별 구성 상이 | ASSISTANT는 RAG/LLM 명시, 기본 모델 이름/버전 미공개. 회사의 SONA/sLLM 보유와 각 고객 추론 공급원은 구별 | [85][66][86] |

## 3.1a 보류 후보

| 후보 | 확인된 것 | 제외/보류 이유 | 근거 |
|---|---|---|---|
| 솔트룩스 | 현재 홈페이지의 루시아 온·에이전트·음성 솔루션 및 일반 기업 AX 사례 | 과거 AICC 경로가 **일반 홈페이지로 리디렉션**. 검색결과의 AICC/Samsung 사례를 제품본문 확인으로 간주할 수 없다. 현행 컨택센터 Copilot의 제품·배포·LLM·PII·가격·한국 실명운영 모두 추가 확인 필요 | [71] |

## 3.2 국내 PII·가격·레퍼런스

| 벤더 | PII 처리: 확인 / 미확인 | 가격모델·숫자 | 한국 레퍼런스·증거 수준 | 생성형/Agentic 검증 한계 | 근거 |
|---|---|---|---|---|---|
| 삼성SDS | PII 전처리·마스킹·가역 금고 상세 미공개 | 모델/숫자 미공개 | 익명 **글로벌 하이테크** 적용(V); 국내 실명 고객은 이 원문 미확인 | 기존 AICC/NLU 운영 ≠ 생성형 Agentic 운영 | [56] |
| LG CNS | 보안/전용 클라우드 설명. 외부 LLM 직전 PII 가명화·토큰복원 구조 미공개 | 구독/사용량 기반 선택, 숫자 미공개. "추가 비용 없이 sLLM 운영"은 무상 전체 서비스 뜻 아님 | **KB금융·현대차·우리은행 상담봇**(V) | 기존 AICC 도입과 최신 RT-Advisor 운영 고객은 구별 | [57][58](시점 미상) |

| 벤더 | PII 처리: 확인 / 미확인 | 가격모델·숫자 | 한국 레퍼런스·증거 수준 | 생성형/Agentic 검증 한계 | 근거 |
|---|---|---|---|---|---|
| KT | 제품별 PII 마스킹·금고·복원 상세 미공개 | 이번 검증 원문에 Assist 숫자 미공개. KT Cloud 가격페이지는 JS 셸만 확보하여 요금표 미확인 | **KT 100번 전체 센터**(V); 케이뱅크/KT엠모바일 믿:음 적용(M); **NH농협은행** 차세대 AICC 2026 완료(M) | 약 400억원은 전체 구축사업 총액, 좌석단가가 아님. M의 성과는 KT 발표 전달 | [59][87][82] |
| SKT | PII 상세 미공개 | 구독형, 금액 미공개 | **SKT 자사 고객센터** 2026 운영(V). 외부 10여 가입 고객은 익명(V) | 자사 사례를 모든 외부 CCaaS 구성에 적용하지 않음 | [61][62] |
| 네이버클라우드 | AiCall 제품의 PII 범위·복원·보존 상세 미확인 | 현재 상품 요금본문 접근 실패, **가격 미확인** | **현대커머셜** 금융 해피콜 2023 운영(C) | 고객 원문은 자동응대 운영 증거. HyperCLOVA X 상담사 Copilot 증거는 아님 | [81] |
| 카카오엔터프라이즈 | 클라우드 녹취보관 설명, PII 전처리/격리 상세 미공개 | 월구독·Basic/Premium 구분(2023); 숫자·현행 요금 미확인 | **브랜디·하우스텝** 운영(V) | 현재 판매 계속 확인 불가. **99.7%는 스마트콜백 포함 상담처리완료율**이지 AI 단독 해결률 아님 | [60] |
| 브리지텍 | 로컬 sLLM 가능. 외부 GPT/Claude 전송 직전 PII 비식별·가역 금고는 미공개 | AoE/Assist 과금모델·숫자 미공개 | **한국증권금융 AI 상담지원**, **NH농협은행 IPCC**, **LH주거복지정보** 구축사례 **목록 제목**(V) | 목록의 IPCC 구축을 AoE Tool/MCP 생성형 운영으로 등치하지 않음. 세부 사례본문 미검증 | [68][69] |
| 넥서스커뮤니티 | PII 상세 미공개 | 합리적 비용/구축 절감 설명뿐, 모델·숫자 미공개 | **S전자·S쇼핑몰·S렌트카** 익명 사례(V) | 국내 CTI 44,000석을 Copilot 좌석 또는 실명 운영 실적으로 쓰지 않음 | [70] |
| 스켈터랩스 | 신한투자증권 사례의 보안 문제 설명은 있으나 PII 기술 흐름 미공개 | 숫자·계약모델 미공개 | **퍼시스·요기요 AICC**, **분당서울대병원 STT**, **신한투자증권 사내 RAG**(V, 목록) | 고객마다 기능이 다름. 병원 STT를 금융 RAG/생성형 음성 Copilot 운영으로 합치지 않음 | [63][64] |
| 마음AI | 레거시 소개에 PII 상세 미공개 | 문의형, AICC 숫자 미공개 | **현대해상 음성봇**, **하나은행 HAI**, **푸본현대생명 TMQA**(V) | 구 마인즈랩/푸터 2022 페이지. 2026 현행 모델·고객 계속운영 미확인 | [65] |

| 벤더 | PII 처리: 확인 / 미확인 | 가격모델·숫자 | 한국 레퍼런스·증거 수준 | 생성형/Agentic 검증 한계 | 근거 |
|---|---|---|---|---|---|
| 페르소나AI | ASSISTANT: **민감정보 자동마스킹**, 그룹별 접근권한·다운로드/호출 로그. KGPT: 개인정보 마스킹·문서보안·학습방지 (별도 제품). 사내 가역 토큰 금고·외부호출 순서·한국어 탐지 실측 미공개 | GEN AICC **사용량 기반**, 숫자 미공개. 이 가격모델을 ASSISTANT 단품에 그대로 적용하지 않음 | 연혁의 **KB금융그룹 8개 계열사 KB AICC**(V); ASSISTANT 페이지의 **한국의료분쟁조정중재원** 사례 링크/표기(V) | 명시된 도입은 있으나 95%/1대 30·생산성 200%↓ 등 정의 불명 벤더 수치를 독립 성능검증으로 쓰지 않음 | [85][66][67] |

## 3.3 해외 3사

| 벤더/제품 | 상담사 지원 vs 자동응대 | 배포/기존시스템 접점 | LLM 공급원 | PII 처리 | 가격 | 한국 레퍼런스 | 원문 |
|---|---|---|---|---|---|---|---|
| **Genesys Cloud Agent Copilot** | 상담사: 지식 자동표시·스크립트/양식·요약·종료코드 추천, 상담사 검토/수정. 자동응대 Virtual Agent 별도 | Cloud 플랫폼 내장 SaaS. BYO/custom connector FAQ를 독립 온프레미스 Copilot 제공으로 해석하지 않음 | 읽은 FAQ에 기본 공급사/모델버전 미확인. BYO/custom integration 지원과 **향후 BYO LLM** 문구를 구별 | 학습 데이터 기부 기본 **옵트아웃**, 동의 데이터는 익명화/검증. 이는 inference 외부전송 전 PII 금고 보장은 아님 | 현행 한국어 CX1/2/3/4 **USD75/115/155/240 per month**, 연간 약정. AI Experience 토큰/추가 사용량 별도 조건. Copilot 단품가격 아님 | **LG U+ 6,000 상담사 Cloud 이전 진행** 발표. 완료·현행 Copilot 실사용은 미확인 | [72][73][74] / 한국사례 [75] |
| **NiCE Copilot for Agents / CXone** | 상담사: 실시간 지식·맥락·next best action·workflow. Automated Summary 단독 선택; Process Automation AI Agents는 add-on. Self-Service AI Agent 별도 | CXone 클라우드 서비스. 기존 NEVA 데스크톱 RPA 설치형 여부를 CXone Copilot 온프레미스 LLM으로 일반화하지 않음 | 페이지에서 CX AI/브랜드 기능 확인. 읽은 제품·거버넌스 문서에 모델 공급사/버전 **미공개**; OpenAI/Azure로 추정하지 않음 | 녹취/전사의 **PII masking/tokenization**, Redaction Profile의 실시간 Copilot 전사 적용. 한국 PII 범위·복원 금고·LLM 전송 전 처리 위치 미확인 | Omnichannel/Essential/Core/Complete **USD110/135/169/209 per agent/month**. Ultimate 표기는 **USD249 / USD0.25**(agent/month / session). 슬래시를 무조건 합산하지 않음. Copilot 단품/한국 견적은 별도 확인 | 국내 실명 **Copilot** 운영 고객 미확인. NICE 녹취/WFM/NEVA 고객은 해당 제품 실적과 별도 | [83][76][77] / 거버넌스 [78] |

| 벤더/제품 | 상담사 지원 vs 자동응대 | 배포/기존시스템 접점 | LLM 공급원 | PII 처리 | 가격 | 한국 레퍼런스 | 원문 |
|---|---|---|---|---|---|---|---|
| **Verint Agent Virtual Assistant / Copilot Bots** | 상담사: Knowledge Automation·Interaction Wrap Up·거래처리 지원. 고객 IVA self-service 별도 | 기존 CX시스템에 통합하는 open approach. 이번 문서에서 신규 Copilot의 완전 온프레미스 추론 보장 미확인 | Da Vinci의 상용/공개/자체 모델 조합. **2026-06 v1.7 공식 PDF**에 OpenAI/GPT·Anthropic Claude·Google Gemini/Gemma·Meta Llama·Qwen 등 지원 모델군 열거. "지원"≠한국 계약의 기본 모델 | PDF의 NER redaction 공급자: **AWS·Microsoft Azure Language·Cobalt·Tonic.AI**. 정확한 적용 순서/한국어 탐지·토큰복원·리전은 미확인 | 숫자 미공개, Contact Sales. 모듈/Bot 구성별 조건 확인 필요 | 국내 실명 **Copilot/생성형 상담사 지원** 고객 미확인. 일반 WFM/녹취 또는 한국 포함 글로벌 매장을 대체 근거로 쓰지 않음 | [79][80][84] |

**가격 비교 주의:** 위 USD 금액은 공개 CCaaS 패키지 표기이며 한국 계약의 세금·전화비·AI 전사/토큰·통합·구축·유지보수 포함 총원가도, 동일 범위 Copilot 단품 비교도 아니다. Genesys 연간 약정/토큰 조건과 NiCE suite·session 조건을 같은 좌석 단가로 놓고 우열을 계산하지 않았다.[73][76]

**문서 날짜 주의:** Verint PDF의 URL 경로에는 /2024/10/이 있지만 실제 읽은 표지 버전은 **Version 1.7 - June 2026**다. 경로 연도를 발행일로 쓰지 않았다.[80]

### 3.4 경쟁의 시사점 [의견]

- FCC RT-Advisor·SONA ASSISTANT·AoE를 직접 비교대상으로 선정. '온프렘+마스킹+요약' 기능만으로 차별화 주장 불가.[57][85][68]
- 사내 가역PII 금고→외부추론 전 보호→사내복원 전구간, 한국어 누출률·개입정확도·총원가·폐쇄망 dependency를 재현가능하게 증명하는 방향 권고. 공개문서 미확인을 경쟁업체 미구현으로 단정 금지.[85][78][80]
- 최신 모델·실명 Agentic 운영·고위험실행 안전성은 고객별RFI 대상. 기존AICC/IPCC/STT 실적·봇 처리율을 Agentic Copilot 실증으로 합치지 않음.

## 4. 기술 아키텍처 선택지

### 4.1 배포경계 비교 [설계 제안]

| 선택지 | 데이터·모델 경계 | 장점 [추정] | 반대조건/비용 | 적합조건 |
|---|---|---|---|---|

| 선택지 | 데이터·모델 경계 | 장점 [추정] | 반대조건/비용 | 적합조건 |
|---|---|---|---|---|
| 완전로컬 | STT/NER/embedding/gate/LLM/로그 모두 고객통제망 | 외부전송 경로를 줄임 | GPU유휴·패치·라이선스·보안운영·동시성 비용. 국산공개가중치도 상용공급권 별도 | 반출제한·현장운영권/장비 예산 확보 |
| 보호된 하이브리드 | 사내수집·PII금고·게이트, 최소화 문맥만 계약된 외부LLM | 모델선택·가변용량 | 가명정보 법적적용·국외이전·감사로그·프롬프트잔여맥락·외부장애 검증 | 반출경로/처리근거/승인 확보 |
| 기존벤더 확장 | 기존AICC/CCaaS·STT/CRM에 모듈 통합 | 재조달·교체범위 감소 가능 | API·SKU·하도급·모델경로 잠김, 기존사업자 통합비 | 이미사용중인제품의지원권 확보 |
| 독립SaaS | 기존텍스트를 SaaS로 전송, 공급자 운영 | 빠른실험 가능 | 금융/공공 조건·민감한 대화·리전/수탁자·로그검증 부담 | 비민감범위·계약/법무승인·고객선택권 확보 |

근거: 배포권·모델카드/라이선스와 금융·개인정보·CSAP 문서. 표의 장단점은 본 연구의 설계 의견이며 개별서비스 허가 아님.[26][8][11]

### 4.1a '프라이빗 Vertex AI/Azure OpenAI가 있는데 왜 못 쓰는가?'

| 선택지 | 확인된 공급자 통제 | 해결하지 못하는 경계 |
|---|---|---|
| Vertex AI 사설 접근 + VPC Service Controls | Private Service Connect/Private Google Access 및 서비스 보호 경계. 사설 통신만으로 공개 API 접근이 제거되는 것은 아니라는 공식 경고.[89] | 공급자 내부 추론, 고객 계약, 잔여 PII, 처리 지역·외부 검색·보존 예외. 온프렘 가중치와 동일하지 않음 |
| Vertex AI 데이터 정책 | 허락 없는 학습 미사용. 남용 감시·grounding·고급 AI 기능·로그 등의 보존 조건/예외 설명.[88] | 무학습≠무전송≠무보존. 개별 서비스/계약과 기관 심의 확인 필요 |
| Azure OpenAI 사설 접근 | Private Endpoint·DNS로 VNet의 사설 IP 연결. 확인 문서는 classic portal 적용 범위.[91] | 사설 엔드포인트만으로 공급자 내부 처리·보존·법률 승인 증명 불가 |
| Azure OpenAI 데이터 정책 | Microsoft가 Azure 환경에서 제공하며 OpenAI 소비자 서비스와 다름. 데이터 비공유·허락 없는 학습 미사용, Global/DataZone 및 남용 감시 저장/검토 구분.[90] | 고객 실제 배포형·국외 접근·상태 저장·modified abuse monitoring 승인 설정은 미확인 |

**[판정]** 두 플랫폼의 기업용 보호 장치가 있다는 사실은 인정. 그 장치만으로 한국 금융·공공 상담 데이터의 모든 반출이 허용된다는 결론도, 보호 장치가 전혀 없다는 결론도 부정확. 구매자는 연결·학습·보존·처리 국가·처리 근거·승인을 따로 검토해야 함. 위 문서 확인일 2026-09-30; Google 구 Vertex URL은 현재 Gemini Enterprise

Agent Platform 문서로 이동했으며 일반 기능을 개별 고객 설정으로 확정하지 않음. (raw/v2-parent-google-governance.md)

## 4.2 추천 경로 [의견]

```
허가된 기존STT API/업무이벤트/상담사 수동호출
  → call/session 소유권·시간·partial/final revision 정합성
  → 사내 정규화·schema 최소화·PII 탐지·불투명토큰+분리금고
  → 로컬 G1: 무개입/규칙/검색/LLM필요/사람검토
  → G2: 승인된 로컬/외부 모델 선택 + 지식문서 접근권 검사
  → 출력근거·token소유권·PII/정책 G3 검사
  → 승인된상담사 UI에만 복원·표시
  → Tool 실행은 별도 deterministic 권한·승인·감사·재시도/롤백
```

- 게이트·embedding·RAG·오류로그도 보호경계에 포함. 외부 게이트가 원문을 매턴 읽으면 비싼LLM 호출감소와 원문노출감소가 분리됨.
- PII 금고는 mapping/키·TTL·테넌트/세션분리·권한·감사·동시성·복구를 갖춘 저장경계. 예제Python dict를 production 금고로 판매하지 않음.[21]
- 조회/권고와 거래·취소·정보변경 쓰기분리; 모델결정 자체를 권한으로 인정하지 않음.

## 4.3 오픈소스 판단·라우팅 비교 — G1/G2/G3 구분

**설계 구분:** G1 = 지금 개입할지(no-call/규칙/검색/상담사 검토), G2 = 생성이 필요할 때 어느 모델인지, G3 = 답변·행동을 허용할지. G1/G2/G3를 하나의 router 성능으로 표현하지 않는다. 이는 연구자의 기능 분해 제안이다.

| 후보 | 사실: 공개/배포 근거 | 주로 해결하는 결정 | 현장 한계·반증점 | 의견: 적합한 사용 |
|---|---|---|---|---|
| TypeSafe Jev | hosted API; 모델 가중치 공개/온프렘 권리 확인 안 됨; adapter MIT.[45][54] | typed 판단 G1/G2/G3를 질문으로 구성 가능.[29] | 한국어 비동등, 네트워크·데이터 경계, 적대 입력, 버전별 calibration.[30][32] | 민감정보 제거 후 외부 benchmark 비교군. 폐쇄망 기본안으로 전제하지 않음 |
| Aurelio semantic-router | 코드 MIT; local HuggingFaceEncoder 지원.[37][18] | utterance 의미 유사도 intent/no-match; G1 rule/RAG/no-call 분기 설계 가능.[18] | 유사도는 교정된 정답 확률 아님. main 1.x 개발/0.x quickstart 혼재. local encoder license·한국어 임계치 별도.[18] | 로컬 다국어 encoder + 규칙의 단순 G1 baseline. 릴리스·모델 고정 |

| 후보 | 사실: 공개/배포 근거 | 주로 해결하는 결정 | 현장 한계·반증점 | 의견: 적합한 사용 |
|---|---|---|---|---|
| RouteLLM | 코드 Apache-2.0; 학습 router·평가 틀 공개.[38][19] | strong/weak 모델 사이 선택 G2.[47] | 기본 mf/sw_ranking은 OpenAI embeddings 요구; local 생성 backend 연결만으로 폐쇄망 아님.[19] | G1 뒤에 선택적으로 평가. BERT/local embedding 조합은 코드·checkpoint·egress 검증 후 채택 |
| vLLM Semantic Router | 코드 Apache-2.0; Envoy·model runtime 사이 control/routing layer.[39][53] | 신호·정책·recipe 기반 모델·실행 경계 선택/조합.[20][53] | Jev와 동등한 단일 작은 판단 모델 아님. classifier/embedding/recipe·backend 통합 및 운영 복잡도 평가 필요.[53] | 다수 모델·위치·정책을 운영해야 하는 단계의 G2/G3 인프라 후보 |
| 별도 소형 한국어 분류기 | 이 행은 제품 사실이 아닌 아키텍처 제안. code/model/data license 각각 확인 필요 | 사람 라벨로 G1의 개입/무개입·abstain 학습 | 이벤트별 라벨·지속 교정 필요. 정확도뿐 아니라 놓친 중요 개입과 nuisance 평가 | 가장 좁은 G1 문제에 우선 비교할 baseline |

RouteLLM 논문 v4(2025-02-23) 본문은 preference data 기반 strong/weak 모델 라우팅이며 한국 컨택센터의 개입 필요성 검증과 다르다.[47] 공개 benchmark 절감 수치가 한국 상담 비용·품질로 재현된다고 주장하지 않는다. 게이트가 모든 턴을 읽으면 호출 수가 줄더라도 원문 노출 면적이 자동으로 줄지 않는다. **추정:** 게이트 자체가 외부에 있다면 보호된 텍스트를 보내는 순서가 중요하다. (raw/tech-routers.md)

## 4.3a 라우터 이외의 로컬 판단·가드레일·PII 후보

**[판정]** JEV-like 대안은 업무 개입 판단, 모델 선택, 콘텐츠 안전, PII 구간 탐지를 나누어 평가. 아래의 한국어·권리·GPU/지연 칸은 공식 문서 확인과 미실측을 구별.

| 후보/정확한 artifact | 기능·한국어 확인 범위 | 코드와 weights·데이터 권리 | GPU·지연 및 채택 조건 |
|---|---|---|---|
| 로컬 sLLM + vLLM v0.15.0 | guided choice에 해당하는 structured_outputs.choice 또는 JSON schema로 허용 라벨/형식을 제한. logprobs로 토큰 점수 반환 가능. 한국어 이해는 선택한 sLLM에 의존한다.[114][115][117] | **엔진 코드 Apache-2.0**. sLLM weights 및 tokenizer/데이터 약관은 별도; 엔진 license로 모델 상업권리를 확정하지 않는다.[116] | GPU 종류·VRAM·동시성·p95 미실측. 모델 크기, 양자화, 입력 길이, 배치/큐에 따라 비용·지연을 따로 시험한다. 형식 일치가 판단의 정답성 또는 교정 확률이라는 결론은 금지. |
| answerdotai/ModernBERT-base | 149M encoder, 카드 언어 en, 영어·코드 사전학습. 업무 분류에는 fine-tuning 필요. 한국어 성능 근거로 GLUE/검색 benchmark를 사용하지 않는다.[118] | weights 카드 **Apache-2.0**. 사용하는 inference 소프트웨어와 fine-tuning 데이터 권리는 별도 확인. 본 카드만으로 모든 학습자료 재사용권리 확정 불가.[118] | GPU/CPU 및 지연 미실측. 한국어 적응·라벨 데이터 필요성부터 평가한다. 영어 카드인데 한국어 완성 분류기로 채택 금지. |

| 후보/정확한 artifact | 기능·한국어 확인 범위 | 코드와 weights·데이터 권리 | GPU·지연 및 채택 조건 |
|---|---|---|---|
| monologg/koelectra-base-v3-discriminator | 카드에 ko 명시, 한국어 사전학습. ELECTRA의 replaced-token discrimination은 업무의도/민감성 분류 라벨이 아니므로 head·fine-tuning 필요.[119] | 저장소 코드 **Apache-2.0**, weights 카드 **Apache-2.0** 각각 확인. 사전학습/현장 fine-tuning 데이터 권리는 별도.[119][120] | GPU/CPU·지연 미실측. 소형 로컬 인코더 분류기의 한국어 후보로 우선 평가하자는 의견. 클래스 불균형과 보류 라벨을 포함한 검증 필요. |
| meta-llama/Llama-Guard-3-8B | prompt/response 콘텐츠 안전 분류, first-token unsafe score. 영어·프랑스어·독일어·힌디어·이탈리아어·포르투갈어·스페인어·태국어 지원; **한국어는 명시 목록 밖**.[121] | 대상 weights **Llama 3.1 Community License**, 단순 Apache/MIT 아님. 수락·표시·허용사용정책 등 약관 확인 필요. 저장소의 다른 LICENSE를 대입하지 않는다.[122][123] | 8B 모델이라는 사실 외 배포 GPU·VRAM·한국어 정확도·지연 미실측. 한국어 안전판정 보조 후보이지 PII 완전탐지기 또는 업무 라우터의 직접 대체가 아님. |
| NVIDIA NeMo Guardrails | input/dialog/retrieval/execution/output rails 및 연결 모델/규칙 제어. rail 없이 설정하면 요청이 LLM으로 전달될 수 있다. 프레임워크 자체가 한국어 모델이 아님.[124] | 라이브러리 코드 **Apache-2.0**. LLM·embedding·checker weights 및 원격 서비스 약관은 각각 별도.[125] | 프레임워크 단독 고정 GPU 요구·지연을 제시할 근거 없음. 연결 모델 호출 수/정책 체인 영향은 미실측. 한국어 정책 문구·우회·tool 입력검증을 추가 시험. |
| GLiNER upstream + urchade/gliner_multi-v2.1 | 일반 NER, multilingual 209M 카드. 한국어 입력 후보이지만 한국어 PII 전용 학습·현장 recall 보장 없음.[127] | upstream 코드 **Apache-2.0**, 해당 weights **Apache-2.0**. 구형 v0/v1 weights의 CC-BY-NC-4.0와 혼동 금지. 모델/데이터 카드와 추가 학습자료 권리 따로 확인.[126][127] | GPU/CPU·VRAM·한국어 PII 누락률·지연 미실측. 숫자식별자 규칙·검사와 병용할 실험 후보라는 의견. |
| GLiNER upstream + urchade/gliner_multi_pii-v1 | PII fine-tune, 선언 언어 **en/fr/de/es/pt/it**에 ko 없음. multilingual 이름만으로 한국어 PII 지원 보장 불가.[128] | upstream 코드 **Apache-2.0**, 이 weights 카드 **Apache-2.0**. synthetic PII 학습 데이터 이름은 명시되지만 데이터별 권리를 여기서 확정하지 않았다.[126][128] | GPU/지연 미실측. 한국어 외삽 후보, 재학습/검증 전 상업 한국어 PII 완성제품으로 분류 금지. |
| GLiNER-ko taeminlee/gliner_ko | ko 명시·한국어 일반 NER 연구목적. 카드의 konne NER 결과는 상담 녹취 PII 누락률이 아니다.[129] | upstream 코드 **Apache-2.0**와 별개로 **weights CC-BY-NC-4.0**. **상업 도입 보류**; 별도 권리확보/상업허용 artifact 선택 필요. 한국어 fork와 mecab-ko 등 의존성 권리도 확인 필요.[126][129] | GPU·지연·현장 한국어 PII 성능 미실측. 연구용 비교 후보에 한정; 라이선스 허가를 받았다고 추정하지 않는다. |

**[설계]** 제약 선택지/JSON으로 출력 형식을 제한해도 정답이나 교정된 업무 확률을 보장하지 않음. 길이가 다른 라벨·제약 전후 점수·top-k·토큰화를 고정하고 별도 한국어 검증셋에서 재현율·점수 교정·판단 보류를 검사. API/모델 실행은 미수행.[114][115][117]

**[권리 유의]** Llama Guard HF 가중치 약관은 3.1, 현재 PurpleLlama 해당 경로 약관은 3.2로 불일치. 실제 artifact·revision·수락 약관을 묶어 확인.[122][123]

**[권리 유의]** 한국어 GLiNER-ko는 CC-BY-NC-4.0, 다언어 PII-v1은 한국어가 명시 지원 목록에 없음. 다언어 일반 GLiNER의 허용형 권리와 한국어 PII 현장 성능은 다른 문제.[127][128][129]

## 4.4 PII 처리와 한국어

- **[사실]** Presidio 한국 정형 ID 탐지기 5종(KrBrn/KrRrn/KrDriverLicense/KrFrn/KrPassport)은 존재하지만 기본 비활성. 언어별 NLP 설정과 이름·주소·STT 숫자 탐지는 별도 검증.[23][22][34]
- **[사실]** 암호화/복호화·사용자 정의 대응표로 복원 가능. mask/redact/hash만으로 자동 복원되는 것은 아님. 공식 mapping 예제는 동시 실행 안전성을 보장하지 않음.[35][21]
- **[추정]** 한글 숫자·띄어쓰기·부분 전사 수정·동명 고객·교차 세션은 미탐/오복원 위험. 출력의 새 위치 정보는 입력과 달라 토큰 소유권 재검사 필요.[55][21]
- **[의견]** 정형 패턴·한국어 NER·정규화·맥락 최소화와 불확실시 차단을 조합. 다른 세션 토큰·변형 토큰 복원 금지. 대응표 손실시 원문을 추정 생성하지 않음.
- **[사실]** Presidio 코드 MIT, Data Privacy Stack 이전 공지 존재. Microsoft 관리형 지원 계약과 동일하지 않음.[36][51]

## 4.5 공개 LLM과 상용 공급권 — 국내 우선·해외 비교

**[확인 범위]** 정확한 10개 ID의 카드·LICENSE 전문/공식 약관을 수신한 비교. 가족명으로 일괄 권리를 판정하지 않으며 최신 전체 모델 목록이 아님. GPU·한국어 상담 품질·지연은 모두 미실측.

'한국어'는 카드·공급자 주장 확인 수준이며 당사 상담성능 실측이 아니다. 배포예제는 공개문서의 존재이지 설치성공/운영SLA 확인이 아니다.

| 정확한 모델 ID | 공식카드 기준 한국어/규모 | 문서상 배포 | LICENSE 전문 확인과 권리/판매 검토 |
|---|---|---|---|
| K-intelligence/Midm-2.0-Base-Instruct | 한국어/영어, 한국 중심 설명, 11.5B dense.[94] | Transformers·vLLM/함수호출 파서 안내.[94] | MIT; 판매/서브라이선스 가능, 저작권·허가고지 유지.[93] |
| K-intelligence/Midm-2.0-Mini-Instruct | 한국어/영어, 2.3B dense; 경량/온디바이스 목적은 공급자 설명.[96] | Transformers·vLLM 예제.[96] | MIT 별도수신; Base에서 유추하지 않음.[95] |

| 정확한 모델 ID | 공식카드 기준 한국어/규모 | 문서상 배포 | LICENSE 전문 확인과 권리/판매 검토 |
|---|---|---|---|
| skt/A.X-4.0 | 한국어 추가학습, 72B 표준.[98] | Transformers·vLLM 도구사용 예제.[98] | Qwen 자체약관, Apache 아님; 고지/수정표시/NOTICE, 상업 제품·서비스 MAU1억 초과 별도허가.[97] |
| skt/A.X-4.0-Light | 한국어 목적, 7B·Light 문맥 16,384.[100] | Transformers·vLLM 도구사용 예제.[100] | Apache-2.0 및 SKT/Qwen NOTICE·상표조건. 표준과 분리.[99] |
| kakaocorp/kanana-1.5-8b-instruct-2505 | 카드 한국어/영어, 8B 계열.[109] | Transformers 예제, 문맥확장 안내.[109] | Apache-2.0 전문; 재배포 고지·수정표시 등.[108] |
| kakaocorp/kanana-1.5-2.1b-instruct-2505 | 카드 한국어/영어, 2.1B 계열.[111] | Transformers 예제.[111] | Apache-2.0 전문 별도수신.[110] |
| kakaocorp/kanana-1.5-15.7b-a3b-instruct | 카드 한국어/영어, MoE; FLOPS 효율은 공급자 주장.[102] | vLLM 예제; MoE 효율을 VRAM/지연 실측으로 전환 금지.[102] | KANANA LICENSE; API/cloud/remote 접근 제공·재판매 및 SI/온프렘/온디바이스 모델 제공·재판매 등 별도상업허가.[101] |
| Qwen/Qwen3-8B | 8.2B, 100+언어·방언 지원 설명; 본 카드만으로 한국 컨택센터 품질 보증 불가.[104] | Transformers·vLLM·SGLang, thinking on/off.[104] | Apache-2.0 전문. 다른 Qwen 크기/세대/배포본 약관으로 확대 금지.[103] |
| google/gemma-3-12b-it | 카드 140+언어 학습 설명; 한국 상담/사투리 성능 별도검증 필요.[105] | 카드 공개수신; HF 가중치 로그인/약관동의 안내, 가중치 수신 미수행.[105] | Gemma Terms, Appendix의 Gemma3 포함; API Hosted Service도 Distribution, 계약·정책 승계 필요.[112][113] |
| naver-hyperclovax/HyperCLOVAX-SEED-Think-14B | 14.74B dense, 한국어/한국문화 평가 공급자 주장.[107] | reasoning 제어·도구사용 및 vLLM 예제.[107] | 전용약관 4조: 계열사 포함 MAU 조건 **또는** NAVER 제품·서비스와 실질적유사/직접경쟁 시 명시허가 필요.[106] |

어느 모델도 이 조사에서 한국 상담 정확도·PII 누출률·tool 성공률·GPU 메모리·TTFT·p95 응답지연·동시성·비용을 측정하지 않았다. '7B라 단일 GPU 가능', 'A3B라 3B 메모리', '한국어 지원이므로 금융상담 적합' 같은 결론을 내리지 않는다. 가중치, KV cache, 문맥, 배치, 양자화, 추론모드가 포함된 재현 가능한 별도 실험이 필요하다(검증 제안/의견).

#### 4.5a 재판매·경쟁 조항

- **HyperCLOVAX 4조:** 계열사 포함 MAU 문턱 **또는** NAVER와 실질적으로 유사/직접 경쟁하는 서비스면 별도 명시 허가 조건. 소기업도 경쟁 조건에서 자동 면제 아님. 상담 Copilot 재판매는 사업구조별 비해당 확인/허가와 법무 검토 전 기본 상용 후보에서 보류. 특정 사업의 위반을 확정하지 않음.[106]
- **A.X:** 표준 72B는 Qwen 자체약관, Light7B는 Apache-2.0. Light의 권리를 표준에 복사 금지.[97][99]
- **Kanana:** Apache 확인은 8b/2.1b -2505 ID. MoE15.7b-a3b는 API 접근 제공/재판매·SI/온프렘·온디바이스 공급의 별도 상업 허가 조건.[101][108][110]
- **Gemma3:** 전용 약관, hosted API도 배포 정의에 포함. Appendix·정책·계약 승계를 확인; 홈페이지 코드 샘플 Apache 표시를 모델 권리로 오해 금지.[112][113]

#### 4.5b 기존 후보 유지·권리 경계

| 특정 ID | 확인 조건 | 사업 판단 |
|---|---|---|
| LGAI-EXAONE/EXAONE-4.0-32B | 1.2-NC, 직접/간접 상업 활용 별도 계약.[25] | 무상 상용 기본 후보 제외 |
| LGAI-EXAONE/K-EXAONE-236B-A23B | 내부 상업 사용과 제3자 상업 공급/서브라이선스 별도 계약 구별.[26][44] | 내부 사용≠고객 설치/호스팅 재판매 |
| upstage/SOLAR-10.7B-v1.0 | base Apache-2.0 카드.[50] | 한국 상담 지시 품질은 미검증 |
| upstage/SOLAR-10.7B-Instruct-v1.0 | instruct CC-BY-NC-4.0.[27] | base 약관을 instruct에 적용 금지 |

**[의견]** 가중치·토크나이저·양자화본·추론 엔진·임베딩·PII·STT의 권리와 구성 목록을 각각 고정. '상용 라이선스가 편함'은 한국어 품질·보안·운영 SLA 우위가 아님. 다운로드 권한/승인·무보증·지원 계약·오프라인 의존성도 확인. 자세한 revision/발행정보: wiki/v2-model-licenses-integrated.

### 4.6 STT·채널·NEVA

| 입력 경로 | 확인 사실 | 현장 반증점 | 추천 순위 [의견] |
|---|---|---|---|
| 기존 STT API/이벤트 | 고객별 API·사용 권한 **미확인** | 통화ID·채널·부분/최종·수정번호·시각이 없으면 부적합 | 권한 확보시 1순위, 중복 전사 방지 |
| 서버별 음성 스트림/SIPREC | RFC는 실시간 RTP·메타데이터와 분리/혼합 구성 모두 포함.[42] | 전환·회의·통화 연결 구간과 고객/상담사 역할 불일치 | 서버 연동권 확보시 PC보다 우선 |
| PC loopback+마이크 | WASAPI loopback은 재생 장치의 혼합음, 프로세스별 수집 샘플 존재.[28][52] | 알림·대기음·에코·회의·VDI·드라이버·장비 변경 | 서버 연동 불가, 소프트폰 음성 경로가 통제될 때 시험 |
| NEVA 화면 값 | 데스크톱 행동·여러 앱 맥락 취득 설명.[48][40] | 임의 STT 화면의 무누락 스트리밍 API·한국 버전·커넥터·SLA **미확인** | API 불가시 지정 필드 읽기 전용 보조 경로 |

출처: Microsoft·IETF·NiCE 공식 문서. 음원 2개≠고객/상담사 역할 2개, 화면 맥락 취득≠전사 이벤트의 계약상 완전성.

- Microsoft 실시간 다채널 문서는 preview/SLA 부재 경고를 포함. 모든 STT의 운영 지원으로 일반화하지 않음.[41]
- NiCE Studio 구판 PDF의 URL2024/03 및 PDF 메타데이터 2022-02-23을 최신 지원 범위로 바꾸어 인용하지 않음.[48]
- **[설계]** 새 발화 의도를 계속 탐지하려면 게이트 앞의 지속 STT 또는 동등 신호가 필요. LLM 선택 호출이 STT 선택 수집과 동일한 것은 아님. TTS는 상담사 화면 보조 MVP의 필수 요소 아님.

## 4.6a STT 상품별 비교

partial은 수정될 수 있는 중간 전사, final은 해당 엔진 계약상 확정 전사를 뜻한다. 스트리밍 연결 수, 물리적 오디오 채널, 음향기반 화자 분리, 고객/상담사 역할 식별은 서로 다른 항목이다. 다음 표의 미확인은 미지원 확정이 아니다.

| 제품/대상 | 한국어·실시간 partial/final | 화자·채널 | 온프렘·license | 가격·품질·GPU/지연 |
|---|---|---|---|---|
| 리턴제로 **VITO(B2C 앱)** | 공식 회사 페이지는 통화녹음 후 문자변환 앱으로 설명. 앱 설명을 B2B streaming API의 partial/final SLA로 대입하지 않는다.[130] | 앱의 화면 화자표시와 B2B API 채널 계약은 별개; 상세 앱 인터페이스 미확인. | 소비자 서비스 이용조건과 기업 온프렘 계약은 별개. 앱 weights 공개 OSS 권리 확인 없음.[130] | B2B API 및 온프렘 가격의 대리변수 아님. 앱 실제 한국어 성능·자원/지연 미실측. |
| 리턴제로 **RTZR STT B2B Cloud API** | 한국어 sommers_ko, whisper의 language=ko, gRPC/WebSocket streaming. 응답 is_final 및 text/start_at/duration/confidence 존재하여 중간·확정을 구분 가능.[131][132] | **일반 파일 API**의 diarization 및 다중채널 확인. 다중채널 때 diarization은 비활성, spk는 채널 ID. streaming 예제는 mono이고 해당 schema로 실시간 화자분리/다채널 동시지원은 확정 불가.[134][135][136] | 상용 서비스. 고객 통합코드와 공급사 모델/서비스 사용권리 별도; 클라우드 API 기능 문서를 기업 온프렘 계약의 전체 명세로 간주하지 않는다.[137] | API 공개 단가 시간당 **1,000/500/400/300원**(월 사용량 누진 구간, VAT 별도), 최소 10초, T3 이상 사전 문의. 다중채널 파일은 채널별 길이 합 과금. 한국어 현장 CER/WER, 동시성, GPU·p95 지연 미실측.[133][136] |
| 리턴제로 **RTZR STT Enterprise On-Prem** | 공식 페이지는 On-Prem 제공 및 customization 안내. Cloud의 모든 streaming 모델/feature가 동일 artifact/SLA로 제공되는지는 견적·명세 확인 필요.[130][137] | 화자/동시접속 관련 기능은 상품 페이지에 있지만 실시간 화자 정확도·물리채널 사양·역할 매핑은 계약별 확인.[137] | 제공 사실 확인. weights 공개 OSS license가 아니라 기업 상용계약 검토 대상.[137] | **On-Prem 가격 숫자 미공개: Contact sales**. GPU/서버 구성·실시간 p95·유지보수 포함가 미확인, 미실측.[137] |

| 제품/대상 | 한국어·실시간 partial/final | 화자·채널 | 온프렘·license | 가격·품질·GPU/지연 |
|---|---|---|---|---|
| **NAVER Cloud CLOVA Speech** | gRPC streaming은 ko/en/ja, 16kHz·1 channel·16bit PCM. text/position·EPD·epFlag/seqId·timestamp/confidence 응답. **RTZR식 `is_final` 계약은 이 문서에서 미확인**; 수정 가능한 partial과 확정 결과를 어떻게 매핑할지 별도 확인. Free 미지원, Basic 전용.[138] | 장문 REST batch에는 diarization과 speakerCountMin/Max 존재. streaming 문서의 mono 입력·응답에 실시간 speaker ID는 확인 못함. batch 화자인식을 streaming으로 확대하지 않는다.[138][139] | 확인한 API는 ncloud endpoint·Classic/VPC용. **VPC ≠ 온프렘**. 공개 weights/OSS license 및 독립 온프렘 배포 권리는 본 조사에서 미확인.[138] | 요금 페이지 검색 단서는 있으나 상품 상세 endpoint는 HTTP 404로 **본문 단가 미확인**. **가격 미공개라고 단정하지 않는다.** Cloud 단가와 별도 기업/on-prem 견적 확인 필요. 현장 한국어 품질·GPU/지연 미실측. raw/v2-tech-clova-price-unconfirmed |
| **OpenAI Whisper 원 구현 + multilingual weights** | tokenizer에 ko/korean. .en과 구분. 기본 transcribe()는 파일 전체를 읽고 30초 sliding window 처리; native live partial/final 서버가 아니다.[140][142] | README의 기본 전사 기능은 화자 ID/고객-상담사 역할 식별을 보장하지 않는다. 채널 분리·음향 화자분리는 별도 설계/검증 필요.[140] | 로컬 self-host 가능. **코드와 원 weights 모두 MIT**라고 공식 README 명시, LICENSE 본문 확인. 입력 음성/학습자료 권리는 별개.[140][141] | OSS license와 무관하게 서버·통합·운영 비용 발생. 공식 VRAM/속도표의 영어/A100 값은 한국어 live E2E p95가 아니므로 이 표에서 수치 대입하지 않음. 실제 GPU/품질/지연 미실측.[140] |
| **SYSTRAN faster-whisper + 지정 weights** | CTranslate2 기반 Whisper 구현, CPU/GPU·양자화 지원. segment generator를 live partial/final 계약으로 간주하지 않는다. ko는 사용하는 multilingual Whisper weights 기준.[142][143] | 별도 integration으로 WhisperX/whisper-diarize 등을 안내하지만 기본 라이브러리의 화자기능이나 모든 구성요소 권리가 자동 확보된 것은 아님.[143] | 코드 **MIT**. 원 Whisper weights MIT 근거는 OpenAI 문서; distil/변환/파생 weights는 선택 artifact별 카드/license 별도 확인. 로컬 배포 후보.[140][144] | API 공급사 공개단가가 없는 self-host 구성; compute/운영은 별도 비용. README benchmark의 특정 음원·정밀도·hardware 결과를 한국어 현장 실시간으로 확대 금지. GPU/지연 미실측.[143] |
| **Whisper 계열 streaming wrapper (UFAL 예시)** | offline 엔진 위에 local agreement 등 확정정책을 추가하는 별도 구성. native Whisper와 구분. README는 2025년부터 SimulStreaming에 대체되어 구형이라고 경고.[145] | 확정정책·화자·역할·중복/취소 semantics를 전체 stack 단위로 확인. | 선택 wrapper·VAD·diarizer·backend weights·데이터 각각 권리 확인 필요. 이 조사는 해당 전체 stack 상업권리를 확정하지 않았다.[145] | 연구 README latency를 한국어 콜센터 실측으로 사용할 수 없음. 최신 stack 선정, 통합 PoC 및 비용 산정 미실행.[145] |

**[품질 판정]** 공급자 한국어 샘플 비교·Whisper 영어/A100·faster-whisper 지정 하드웨어 결과로 현장 우승자 선정 불가. 동일 승인 음원의 오류율(CER/WER), 숫자·금액·이름·주소 일치, 부분 결과 뒤집힘, 확정 후 수정, 누락·중복·역할·동시성별 p50/p95를 측정할 것. 실제 음성/모델 호출은 미실행.[130][140][143]

**[데이터 경계]** Cloud STT에 원음이 먼저 전달되면 텍스트를 뒤에서 로컬 마스킹해도 음성 반출이 사라지지 않음. 음성 처리·보존·재위탁·국외 이전의 승인 경로가 없으면 온프렘 또는 이미 승인된 STT 경로를 선택한다는 설계 의견. wiki/v2-gating-stt-integrated

### 4.7 비용·지연 검증식 [가정·설계]

C_gate = 평가 횟수×게이트 비용 + 생성 횟수×생성 비용 + STT + 금고 + 인프라 + 운영 + 재작업

동일 품질·동시성·감사 범위의 기준 구성과 비교. 생성 부분만 보면 g+rL<L, 즉 r<1-g/L일 때 이득.

- 산술 가정:1만회, 게이트 USD0.00002/회, 생성 USD0.001/회, 생성 승격률 25% → 무게이트 USD10, 게이트 USD2.70, 생성 부분 73% 감소. **실측 가격·총원가 절감률 아님.** wiki/v2-calculations
- 저렴한 생성 모델·GPU 유휴·높은 승격률이면 경제성 소멸 가능. 게이트가 추가되어 개입 요청의 지연은 증가할 수도 있음.
- 수집→STT→PII→판단→검색→생성→화면 전체 p50/p95/p99·시간 초과·누락·중복을 측정. 구간 p95를 단순 합산해 전체 p95로 표시하지 않음.
- 동일 통화를 전량 호출·규칙만·로컬 분류기·semantic-router로 비교. 실제 청구·GPU 사용 시간·재작업 포함.

## 5. 규제·보안 요건 — 일반 법문과 개별 승인 분리

**법률 자문 아님.** 개인정보보호법·신용정보법 확인판 2026-09-11 시행, 전자금융감독규정 확인판 2026-07-15 시행. 개별 고객·상품·배포·데이터별 법무/보안 판단 필요.[4][5][6]

### 5.1 개인정보·금융·공공 경계

| 쟁점 | 확인된 사실·날짜 | 설계·진입 판단 [추정/의견] | 미확인·반대 조건 |
|---|---|---|---|
| 가명 처리 | 가명정보도 개인정보. 제28조의2 특례는 통계·연구·공익 기록 보존 등.[4] | 토큰 금고는 노출 감소 통제. 고객별 복원 상담과 연구 특례 구별 | 처리 근거·목적·복원 적법성 |
| 개인신용정보 | 제공·이용 목적, 위탁 예외, 가명정보 분리 규칙 존재.[5] | 잔액·거래·연체 정보와 위탁 역할 검토 | 외부 호출을 모두 제3자 제공 또는 모두 위탁으로 단정 금지 |
| 국외 이전 | 조회·위탁·보관 포함; 여러 법정 처리 근거 존재.[4] | 사설 연결·국내 리전만으로 해외 열람·백업·로그 경로가 없다고 단정 금지 | 실제 국가·수탁자·지원 인력·보존 설정 |
| 금융 클라우드 | 중요도·사업자·안전조치·위원회 절차 및 국내 설치 조건/예외 존재.[6] | 클라우드 허용과 상담 프롬프트 반출 허용은 별개 | 데이터·기관·용도별 심의 |

| 쟁점 | 확인된 사실·날짜 | 설계·진입 판단 [추정/의견] | 미확인·반대 조건 |
|---|---|---|---|
| SaaS 망분리 예외 | 2026-04-20 시행; 고유식별·개인신용정보 제외, 가명정보는 별도 혁신금융 지정 절차.[8][15] | 전면 허용·전면 금지 모두 부정확 | 시행세칙 별표7 전체 대체 통제 대조 미확인 |
| 생성형 AI 혁신금융 | 금융위 2026-04-15: 지정 169건, 출시 45개. 기존 모델 변경 절차 간소화.[7] | 승인 제약이 남고 사용 경로도 존재 | 상담 용도 지정/출시 수 미확인. 45/169를 출시율로 계산 금지 |
| 9월 긴급 완화 | 2026-09-03 발표는 보안 취약점 탐지·개선 시험과 10월 계획.[9] | 상담 Copilot 포괄 허용으로 확대 금지 | 10월 선정은 기준일 이후 |
| CSAP | KISA2026.07 FAQ: LLM API 연계 자체 제한 없음, 데이터별 평가·전용 구성 인증 범위 구분.[11] | 기반 IaaS 인증이 신규 SaaS 인증을 자동 보증하지 않음 | 등급·구성·기관 보안성 검토 |
| 공공 데이터 경계 | 해설서 14.2.1 국내 위치·영역 분리, FAQ 내부 업무 보안성 검토.[12][11] | 배포·운영 인력·백업 경로 증빙 | 기관별 기준·예외 적용 |
| 공공 AI 방향 | 행안부 2026-06-10 공동 기반·중복 GPU 감소·단계적 RAG 방향.[14] | 기관별 신규 온프렘만을 시장 전제로 삼지 않음 | 개별 조달·공동 기반 예외 조건 |

출처: 각 행 법령·금융위·KISA·행안부. wiki/regulation-draft에서 원문 버전·발행일 확인.

**개인정보위 유사 사례 [사실·범위 교정]:** 개인정보위 2025.8 안내서 인쇄 20쪽의 통화 녹음·요약 LLM 사례는 **소비자 서비스의 국외 이전 고지 문제를 다룬 유사 사례**. 한국 B2B 컨택센터 사업자가 동일 구성으로 제재받은 직접 실증으로 사용하지 않음. 처리 국가·고지·처리 근거를 별도 검토해야 한다는 참고 근거.[16]

## 5.2 AI기본법·통화 수집·상담사 개인정보

### 5.2a AI기본법 — 해당성·투명성·책임

- **법문 버전:** 최초 시행 2026-01-22와 확인 현행 법률 2026-07-21, 시행령 2026-08-20을 구분. 미래 입법예고를 현행 의무로 채택하지 않음.[146][147]
- **고영향 검토:** 중대 위험/기본권 영향과 열거 활용 영역을 함께 검토. 법제33조 사전 해당성 검토. 대출·채용·공공 자격 등의 실질 판단, 차량 주요 작동·안전 안내와 일반 상품 FAQ를 분리. 사람의 승인 버튼만으로 비고영향 확정 불가.[146][147]
- **생성형 투명성:** 법제31조 고지·표시를 고영향 여부와 별도로 검토. 시행령제23조의 내부업무 전용 등 전부/일부 예외가 있으나 외부 공급자와 고객 사업자·이용자·최종 전달을 구분. '백엔드이므로 면제' 또는 '모든 전화에 동일 AI 동의 의무'라는 단정 금지.[146][147]
- **책임:** 개발/이용사업자 지위를 공급구조별 검토. 고영향이면 위험 관리·설명·사람 감독·문서 등 제34조 검토, 시행령제27조 근거문서 5년 보관. 민간 영향 평가는 노력 의무이며 모든 Copilot의 일률적 사전허가라고 표현하지 않음.[146][147]

- **자동화된 결정:** 개인정보법제37조의2는 완전히 자동화된 개인정보 처리 결정이라는 전제. 실질적 검토가 있는 권고와 형식 승인/자동 신용·인사결정을 구분. 동의/법령/계약 근거의 거부권 단서가 설명·공개 등을 전부 면제하지는 않음.[149][150]

**5.2b 통신비밀보호법 — 루프백 당사자·제3자 구조**

- 법제2·3·14조의 감청/비공개 타인 간 대화 금지를 실제 통화 당사자·수집·저장·전송 주체와 대조. 'PC에 설치했으니 언제나 당사자 녹음'이라는 결론 금지.[148]
- 대법원 2002-10-08 2002도123은 당사자 일방의 자기 통화 녹음과 제3자의 일방 동의 녹음을 구분. 대법원 2006-10-12 2006도4981은 대화 참여자의 녹음 사안. 외부 AI업체가 자동으로 통화 당사자가 되는 판례가 아님.[156][155]
- **고지≠동의.** 당사자 자기 통화에 통신비밀보호법상 항상 상대방 동의가 필요하다고 쓰지 않음. 반대로 제3자 감청 구조에서 "녹음됩니다" 안내만으로 유효한 동의가 된다고 단정 금지. 후속 위탁·학습·반출의 개인정보 근거도 별도.[148][156][149]

**5.2c 상담사 개인정보·QA·노사 절차**

- 고객 정보와 상담사 음성·이름·로그·생산성/감정 점수 모두 목적별 근거·최소 수집·보존·접근을 검토. 법제15·18조, QA 추가 이용의 관련성/예측 가능성/불이익/안전조치, 학습 목적, 위탁/재위탁·국외 이전을 구분. 계약 또는 녹음 고지가 무제한 모니터링·학습 권리가 아님.[149][150]
- 근로자참여법제4조 사업장 규모·설치 의무를 확인하고, 제20조 협의사항의 신기술·근로자 감시 설비, 제21조 교육/능력개발 기본계획 의결을 구분. 모든 코칭/신기술에 무조건 노조 동의라는 뜻은 아님.[151]
- 근로기준법제93·94조 취업규칙의 의견 청취·불이익 변경 동의를 실제 평가·임금·징계에 연결해 검토. Copilot 명칭만으로 감시 설비성·불이익 여부를 확정하지 않음.[152]
- **법무 검토 전제:** 실제 기능·노사협약·취업규칙·처리 목적은 미확인. 초기 권고 전용·최소 수집도 자동 면제 아님. wiki/v2-regulation-employment-integrated

**납품 체크 [제안]:** 데이터 흐름·위수탁/재위탁·국외 이전 → 역할·세션 경계 → 보존·키·로그 → PII/프롬프트 주입·권한 우회 시험 → 변경 관리·구성 목록 → 장애·긴급 중지 → 고객 법무·보안·노사·조달 승인. 학습 미사용을 적법성 또는 무누출 보장과 혼동 금지.

## 5a. JEV·LAYA — 무엇이고 왜 지금인가

---

§6의 가설③(JEV 같은 빠른 판단 모델로 필요할 때만 LLM 호출)을 판정하기 전에, JEV와 LAYA가 무엇인지 비전문가 기준으로 먼저 정리한다. 표기: **[사실]** 원출처에서 확인한 설명·기록(성능을 우리가 재현했다는 뜻 아님),

**[발행자 실험]** 공급자 측 측정, **[독립 실험]** 외부 작성자 자신의 시험, **[의견]** 연구자 해석, **[설계]** 아직 실행하지 않은 적용 예. 확인 기준일 2026-09-30.

## 5a.1 JEV란 — 글을 쓰는 작가가 아니라, 다음 갈림길을 고르는 판단 부품

- **[사실] 정의:** JEV(공식 표기 Jev)는 TypeSafe AI의 System One 모델. 현재 상태와 미리 정한 질문을 보내면 **선택(Choice)·점수(Score)·참일 확률(Noul, 0~1)**을 돌려주는 비생성형 판단 모델이며, 자유로운 답변 문장을 쓰지 않는다. 고른 값에 따른 실제 실행은 호출한 프로그램이 따로 한다.[29]
- **[의견] 비유:** 생성형 LLM이 '상담 내용을 읽고 후처리 문장을 쓰는 작가'라면, JEV는 '지금 작가를 부를지, 자료를 더 볼지, 사람에게 넘길지 고르는 판단 부품'. 작가를 대체하는 것이 아니라 **작가를 부르는 시점을 보강**하는 부품.
- **[설계] 입력→출력 예시(API 실제 응답 아님):** 상태 상담 종료 의사 확인 / 처리 내용 확인 / 후처리 초안 없음 + 질문 '지금 가장 알맞은 행동은?' → 추가확인 / 초안작성요청 / 사람확인 / 보류 중 하나와 선택지별 확률. 같은 상태에 '초안 작성에 필요한 정보가 얼마나 갖춰졌나'(점수), '고객이 상담 종료를 명시했나'(참일 확률)를 따로 물을 수 있다. 출력값·확률·성공률은 만들어 제시하지 않는다.[29]
- **[사실] 가격:** jev-1.13.0 공식 요금표 기준 **입력 100만 토큰당 USD 0.042, 출력 토큰 무료**(2026-09-30 재조회). API 입력비만의 가격이며 글을 쓰는 LLM·STT·서버·운영 인건비는 별도.[30]
- **[계산 가정]** 판단 1회에 입력 1,000토큰, 10만 회라면 1,000 × 100,000 ÷ 1,000,000 × USD 0.042 = USD 4.20. 질문 수·문맥 길이·재시도에 따라 달라지는 API 입력비 예시.
- **[발행자 실험] 속도:** 2026-09-15 발표의 응답시간 **70~500ms**는 발행자가 미국 서부 노트북에서 서비스를 호출해 잰 값. 한국망 SLA나 상담 후처리 전체 시간이 아님.[191]
- **[독립 실험]** Construct는 약 400토큰·질문 1개 **358~611ms**, 같은 상태의 질문 7개 377~544ms를 보고(2026-09-20, jev-1.13.0, Cloudflare AI Gateway 경유, 저자 측 시험·재현 안 함).[184]
- **핵심:** 답변을 쓰는 AI가 아니라 제한된 선택을 값으로 돌려주는 AI. 형식에 맞는 값이 곧 맞는 판단은 아니며, 0~1 확률도 현장에서 교정된 확률이라는 보증은 없다.[29][188]

## 5a.2 왜 지금 뜨거운가 — 에이전트의 반복 판단을 값싸고 빠르게 떼어 낼 수 있다는 기대

| 날짜 | 확인된 사건 | 읽어야 할 범위 |
|---|---|---|
| 2026-09-15 | TypeSafe, System One 모델·JEV early access 발표 | 모델 서비스 발표일. 대기업 업무 실증은 아님.[191] |
| 2026-09-16 | Vercel AI Gateway, JEV 제공 발표 · browser-use/jev-ultrafast 저장소 생성(UTC, 한국 시각 09-17) | 기존 개발 경로에 판단 API가 들어온 사건. 저장소 생성 시각은 공개·배포 시각과 다를 수 있음.[182][179] |

| 날짜 | 확인된 사건 | 읽어야 할 범위 |
|---|---|---|
| 2026-09-18 | Vercel, 출시 24시간 이용 확산 발표 · LAYA GitHub 저장소 생성 | 서로 다른 사건. LAYA 저장소 생성일을 공식 모델 출시일로 보지 않음.[181][185] |
| 2026-09-21 | Simon Willison의 해설·우려, Anthus의 JEV/LAYA 직접 비교 게시 | 환호만이 아니라 평가·설명 가능성·대안 비교로 논의 확장.[189][186] |
| 2026-09-26~27 | 초기 증거 감사·확률 일관성 분석 논문(arXiv 초록) | 초록까지만 확인. 학계 합의·완결된 검증 아님.[187][188] |
| 2026-09-29~30 | Qualiteg 직접 API 실험, Construct 운영 통합 사례 게시 | 작성자 자신의 호출·통합 기록. 국내 상담 현장 성공 증거 아님.[183][184] |

- **[플랫폼 발표] 확산 속도:** Vercel은 출시 24시간 만에 **AI Gateway 유료 팀의 거의 13%가 JEV를 사용**했고, 사용 팀 수가 이전 모델 출시 때의 **두 배 이상**이라고 발표. 해당 플랫폼·기간·유료 팀 모집단 한정 수치이며 전체 AI 시장 점유율·JEV 유료 계약 수·지속 이용률로 확대하지 않는다.[181]
- **[사실] 관심 지표:** browser-use/jev-ultrafast(Browser Use의 JEV 기반 브라우저 에이전트, MIT) GitHub 별 **약 2.15만**, NandhaKishorM/laya(Apache-2.0) 별 **약 2.9만**(모두 2026-09-30 기준). 별은 관심·북마크 지표이지 설치 수·활성 사용자·구매 수가 아님.[179][185]
- **[개발자 의견]** Simon Willison은 'System One'보다 **decision models**라는 이름을 선호하며 분류·우선순위·검색 재정렬 용도에 관심을 보였고, 설명 없는 숫자 출력의 블랙박스·편향 위험과 더 엄격한 평가 필요성을 함께 제기. 저자 관점이지 업계 합의 아님.[189]
- **[의견] 왜 판단 계층을 떼어 내려 하나:** ① 비용 — 글이 필요 없는 판단까지 LLM이 문장을 생성하면 판단용 생성 비용이 반복된다. 다만 저가 생성형 모델·규칙·기존 분류기로 충분하면 추가 계층이 불필요할 수 있다. ② 지연 — 에이전트는 관찰 → 행동 선택 → 실행 → 결과 관찰을 반복하므로 작은 판단 지연도 누적된다. 병목이 STT·화면·업무 API라면 JEV만 빨라져도 전체는 크게 빨라지지 않는다. Browser Use 공개 코드도 화면 호출 최적화와 글자 입력용 작은 LLM을 함께 쓴다.[190] ③ 에이전트 확산 — 위 초기 지표·공개 코드·운영 통합은 기존 에이전트에 제한된 판단 부품을 붙이려는 움직임의 증거. 국내 구매 수요의 증거는 아니다.
- **[반대 증거]** 초기 감사 논문 초록은 속도·비용 장점과 함께 **어려운 과제에서 정확도 격차**를, 다른 논문 초록은 질문 사이 **확률 일관성 위반**을 보고.[187][188] Qualiteg의 작은 일본어 시험은 가능성과 실패를 함께 보였고, Anthus 비교에서는 기본 LAYA보다 JEV가 나았다가 **추가 학습한 LAYA에서 순위가 바뀌었다**.[183][186] '모든 판단에서 LLM보다 정확하다', '일본어 결과가 한국어에서도 같다'는 주장은 근거 부족.
- **핵심:** 지금의 검증 질문은 'AI가 필요한가'가 아니라 **'이미 쓰는 AI에서 매번 글을 생성할 필요가 없는 판단을 떼어 낼 수 있나'**. 규칙·기존 분류기 대비 품질을 유지하면서 전체 원가·전 구간 지연을 낮추지 못하면 보류.

## 5a.3 LAYA란 — JEV와 무엇이 다른가

- **[사실] 정체:** LAYA는 ConvAI Innovations(GitHub NandhaKishorM)가 코드와 체크포인트를 공개한 **별도의 비생성형 판단 모델·런타임**. 비슷한 형식의 판단을 사내 설치(온프렘)로 수행하도록 공개됐으나, **JEV 공식 가중치 공개판이나 동일 모델이라는 근거는 없다**.[164][157]
- **[의견] 가설 원문과의 관계:** 대표님 가설③의 '온프렘에 올리려면 오픈소스여야 한다'는 요구에 맞는 후보는 JEV 자체가 아니라 LAYA 같은 별도 오픈 모델. LAYA도 JEV와 같은 세 가지 판단값(선택·점수·참일 확률)을 돌려주며, 자체 README가 Jev API와 응답 스키마가 같다고 밝힌다(발행자 주장). 다만 신뢰도(confidence) 계산식이 JEV와 달라 JEV에서 정한 기준값을 그대로 옮겨 쓸 수 없고, 선택지가 약 20개를 넘으면 잘린다.[157]

| 항목 | JEV | LAYA |
|---|---|---|
| 만든 곳·제공 | TypeSafe AI · 표준 약관의 호스팅 API. 기업별 온프렘 공급은 미확인.[30][45] | ConvAI Innovations · Python 코드·학습/교정·서빙 구성과 모델 파일 공개.[157][164] |
| 권리 | 서비스·출력을 이용한 증류·모방·유사 경쟁 개발 금지 표준 조항.[45] | 코드 LICENSE 전문 Apache-2.0, HF 가중치 카드 Apache-2.0 선언. 제3자 백본·데이터·의존성 권리는 별도 확인.[158][164] |
| 비용 | 입력 100만 토큰당 USD 0.042, 출력 무료.[30] | 토큰 API 요금은 없으나 CPU/GPU·운영·모니터링·학습 원가 발생 |
| 한국어 | 영어 위주 학습, 다른 언어 동등 정확도 미보장.[30] | 영어 기본판(421M)이 아닌 **다국어 체크포인트(322M)**가 한국어 후보. 한국어 상담 품질 미실측.[164] |
| 공개 시점 | 2026-09-15 발표.[191] | GitHub 저장소 2026-09-18 생성.[185] |

- **[발행자 실험] 한국어 성능 근거:** 발행자 원결과(2026-09-24, SDK 0.3.20, Intel CPU/Windows, MASSIVE 테스트 한국어 100문항·20선택지)에서 정확도 **영어 기본판 0.110 → 다국어판 0.470**, 확률 교정 오차(ECE, 낮을수록 좋음) **0.6804 → 0.3108**.[165]
- **한계:** 발행자 실험이며 독립 재현 없음. 100문항·일반 의도 분류 과제이고 **자동차 상담·개인정보 탐지·Agentic 성공률이 아니다**. 다국어판 0.47도 실무 수준이라는 뜻이 아니다. 발행자의 대표 속도 32.8ms(T4·질문 조건)와 JEV 236~276ms는 서로 다른 실험 수치라 '더 빠르고 더 정확하다'로 결론 내지 않는다.[159]
- **[독립 실험] 직접 비교:** Anthus는 같은 피드백 140라벨 과제에서 JEV 87%, 기본 LAYA 80%, 가중치를 추가 학습한 LAYA 90%를 보고. 도메인·학습·교정 조건에 따라 순위가 바뀐다는 뜻이며 우리 업무 성능이 아니다.[186]
- **핵심:** 온프렘이 필요하면 LAYA 다국어판이 후보. 채택은 한국어 상담 데이터에서 규칙·기존 분류기와 같은 조건으로 비교한 뒤.

## 5a.4 무엇을 할 수 있나 — 공개 사례 5개와 컨택센터 적용 설계

| 용도 | 공개 구현·측정 | 증거 수준과 한계 |
|---|---|---|
| ① 브라우저 다음 동작·대상 선택 | Browser Use: 관찰한 화면 요소에서 클릭·입력·대기 등과 대상을 고르고, 입력 문자는 작은 LLM에 요청. Google Flights 데모 **7.073초** | 공개 코드·발행자 실험, 1과제. 최초 탐색·브라우저 설치/연결·최종 검증은 계측 밖. 항공권 구매 아님.[180][190] |
| ② 문의 분류·배정 | Qualiteg: 일본어 지원 티켓을 청구/기술/계정/영업/기타로 배정, **29/30 정답** | 독립 실험. 저자가 만든 짧은 티켓·정답표, 한국어 상담 성능 아님.[183] |
| ③ LLM 앞단 가드레일 판단 | Qualiteg: 프롬프트 인젝션 여부·개인정보 포함 여부를 참일 확률로 판정. 첫 시험 개인정보 **23/24**, 질문 보완 뒤 같은 세트 **24/24** | 독립 실험. 계좌 관련 누락 1건, 같은 세트 재사용. 개인정보 위치 탐지·치환까지 하는 증거 아님.[183] |
| ④ 위험 행동의 사람 확인 추천 | Qualiteg: 명령 위험도 점수와 사람 승인 필요 여부. 위험 등급 **21/24**, 사람 확인 신호 **11/12** | 독립 실험. 모델 추천이 실제 권한·정책 엔진을 대체하면 안 됨.[183] |
| ⑤ 에이전트 기억 속 동일 대상 판별 | Construct: 이름·별칭 후보가 같은 사람·프로젝트인지 고르고 문턱을 넘으면 병합. 최대 세 번의 순차 LLM 판정 호출을 한 번으로 대체했다고 보고 | 운영 통합에 대한 저자 발표. 운영 로그·성과는 재현 안 함.[184] |

이 표에 국내 고객센터 상용 운영 실증은 없다. 자체 시험(비공개)에서도 브라우저 작업 성공·실패 기록이 함께 있었으며, 실제 상담 후처리 성공·운영 통합·보안 승인 근거로 쓰지 않는다.

**[설계] 컨택센터 적용 — 실시간 대화 → 빠른 판단 → 필요할 때만 비싼 LLM 호출 (실측·고객 승인 전)**

| 판단할 일 | JEV(또는 LAYA 동등 판단)에 맡길 제한된 질문 | 다음 동작과 필수 통제 |
|---|---|---|
| 후처리 초안 필요 여부 | 선택: 계속상담 / 추가확인 / 초안요청 / 사람확인 | 초안이 필요할 때만 승인된 LLM 호출. 요약 문장 생성은 판단 모델이 하지 않음 |
| 상담 단계 판정 | 선택: 본인확인 / 요구파악 / 안내 / 종료확인 / 불명확 | 가이드 화면 전환 후보 제시. 단계 판정으로 실제 본인인증·접근권한을 주지 않음 |
| 에스컬레이션 | 점수: 불만·위험·불확실성, 참/거짓 확률: 관리자 검토 필요 여부 | 관리자 검토 추천. 문턱은 업무 정책·오탐/누락 비용에 따라 승인 |
| LLM 호출 게이팅 | 참/거짓 확률: 새 질문·상태 변화가 있는가, 선택: 규칙처리 / 지식검색 / 초안작성 / 보류 | 이전 답 재사용·규칙/검색으로 해결되는 건은 LLM을 부르지 않음. 이름·주소 미탐을 해결한다고 가정하지 않음 |

- **[설계] 흐름:** 승인된 STT/화면 정보 → 사내 PII 탐지·토큰화 → 빠른 판단(사내 LAYA 후보 또는 승인된 JEV 경로) → 결정적 정책·권한 확인 → 필요할 때만 승인된 LLM 호출 → 출력·도구·기록 검사 → 상담사 승인 →

허용된 RPA 실행 → 결과 확인·재계획 또는 보류. **JEV API는 외부 경로이므로 PII 원문을 보낸 뒤 안전 여부를 판단받는 구조는 제안하지 않는다.**

- **대표 제공 현장 정보 — 공식 확인 전:** PoC 후보인 현대차 고객센터는 알피에이(주) NICE RPA 기반 실시간 상담가이드가 736석에 설치된 현장으로, 업무·전산 흐름에 대한 이해가 진입 강점. 이 현장에서는 HChat(Vertex AI·Azure OpenAI 경유로 Anthropic·OpenAI·Gemini 저가 모델 사용 가능)을 쓸 수 있고 opencode 사용이 허용되어 있으며, 「고객센터는 개인정보 때문에 AI 연계 후처리 자동화가 쉽지 않다」고 들었다. 위 흐름의 '승인된 LLM 호출' 자리에 HChat을 두는 안이 1순위 후보이나, HChat이 기존에 승인됐다는 사실만으로 새 상담정보·토큰·판단 모델·도구까지 승인됐다고 보지 않는다.
- **[설계] 채택·중단 조건(연구자 제안, 법규·업계 표준 아님):** 같은 한국어 홀드아웃에서 규칙·기존 분류기와 **초안 필요 누락·불필요 호출·보류율·p95 전 구간 지연·총원가**를 비교. 더 빠르더라도 위험 누락이 허용 범위를 넘으면 보류, 승인되지 않은 원문 송신·타 세션 복원이 1건이라도 나오면 중단.
- **[의견] Agentic 여부는 별도:** 판단 모델이 도구 이름을 골랐다는 사실만으로 Agentic Copilot 실증이 아니다. 말(생성)·허용 도구 실행·결과를 읽고 다음 행동 선택/재계획을 수행한 기록과 승인·보류 통제를 확인해야 한다.
- **핵심:** 당장 검증할 것은 **'후처리 초안을 정말 필요할 때만 만들게 하는 감독형 판단 계층'**. 전체 상담 자동화·LLM 대체·보안 보증을 약속하지 않는다.

## 5b. 두 가지 핵심 질문 — 정말 LLM인가, 진짜 Agentic인가

---

§6 가설 판정 전에 대표님이 가장 중요하게 본 추가 질문 두 개를 원문 → 증거 → 판정 순서로 정리한다. 표의 **확인**은 공개 문서의 설명을 확인했다는 뜻이며, 모델 호출을 재현하거나 고객 설치판을 실사한 것이 아니다. 확인 기준일 2026-09-30.

### 5b.1 질문①「AI Copilot, 정말 LLM인가?」

**질문 원문:** 진짜 AI Copilot이라고 하는데, LLM이 사용된게 맞을까? 특히 KT도? 그냥 룰베이스 맥락분석으로 하는 거에 AI만 붙인게 아닐까?

**판정:** KT는 룰에 AI 이름만 붙인 게 아니다 — KT 자체 기술문서로 상담 지식추천·요약에 LLM 사용이 확인된다(모델·버전 미공개). 다른 국내 Copilot은 '공식 제품 설명상 LLM 주장' 수준이고, 실제 모델 실행을 독립 확인한 곳은 없다.[192][193]

**증거 등급 [연구자 기준, 공인 인증 아님]:** ① 자체 기술문서·공시 확인 > ② 공식 제품 주장 > ③ 보도 > ④ 미확인·룰/NLU로 보임. 제품 소개 페이지의 LLM 문구는 ②(주장)이지 구현 확인이 아니다. 규칙·NLU·분류기도 넓은 의미의 AI이므로 **'LLM 근거 미확보'는 'AI가 아님·허위'와 다르다**.

| 벤더·제품 | 등급 | 근거 | 미확인·주의 |
|---|---|---|---|
| **KT 상담Assist — 지식추천·상담 요약** | **① 확인** | KT 기술블로그(2025-10-29):「LLM과 RAG 기술을 활용하여 지식추천 Agent를 개발」, 2025년 3월부터 전국 KT 고객센터 사용, 1·2차 파일럿은 AWS Bedrock·MS Azure.[192] KT ESG REPORT 2025(문서 69쪽):「2024년부터 실시간 상담사 지원 부분에 MS-Azure, 챗GPT, LLM엔진을 결합」,「LLM 기술 덕분에 상담 요약의 정확도와 속도도 개선」.[193] | 모델명·버전 미공개. B2B 고객 설치판이 같은 구성인지 미확인. 2024~2025 서술을 2026 전 상품에 일반화하지 않음 |
| KT 그룹사 AICC(케이뱅크 등) | ③ 보도 | 믿:음 적용 발표 전달 기사(2025-07-30).[87] | 상담Assist의 실제 모델이 믿:음이라고 쓰지 않음 — 모델명·버전은 미공개 |
| LG CNS FCC RT-Advisor | ② 공식 제품 주장 | 상담 전용 sLLM·RAG 명시.[57] | 정확 모델·고객별 실행 엔진 미확인 |
| 브리지텍 AoE | ② 공식 제품 주장 | 자체 sLLM + GPT·Claude 연동 명시 — 국내 중 가장 구체적.[68] | 어느 기능·고객이 어느 모델로 운영되는지 미확인. 연동 가능 ≠ 실제 사용 |
| 페르소나AI SONA ASSISTANT | ② 공식 제품 주장 | 「RAG/LLM이 필요한 기능에 실시간으로 동작」.[85] | 기본 모델 미공개 |
| NiCE·Genesys Copilot(해외 비교) | ② 공식 제품 주장 | 생성 요약·초안 기능 공개.[83][199][72][74] | 모델 공급원 미확인 |
| SKT AI CCaaS · 삼성SDS AICC · 카카오 i 커넥트센터 · 네이버 CLOVA AiCall · 솔트룩스 AICC | ④ 미확인 | 확보 자료에 해당 기능의 생성형 LLM 근거 없음.[61][62][56][60][81] 솔트룩스는 제품 본문 미확보 | 카카오·네이버는 상담사 Copilot이 아닌 자동응대 사례. 다른 LLM 보유를 해당 기능으로 전이하지 않음 |
| 마음AI FAST AICC(구 페이지) | ④ 룰·NLU로 보임 | 패턴분류·시나리오형 챗봇 중심 설명.[65] | 확보한 구 페이지 범위만. 최신 마음AI 전체 판정 아님 |

- **판정 변경 조건 [제안]:** 도입 검토 시 기능별 모델 ID·버전·배포도, 개인정보를 제거한 요청/응답 기록, 규칙 대체 경로 기록을 요구. 연결 증거가 나오면 ②→①로 상향, 해당 설치판이 규칙·추출식만 쓴다는 설정·실행 증거가 나오면 그 설치판에 한해 ④(룰·NLU)로 변경. 비공개라는 이유만으로 미사용으로 확정하지 않음.
- **핵심:** 'LLM인가'는 회사 이름이 아니라 **기능·설치판별로** 확인할 질문. KT 일부 상담 지원 기능의 LLM 사용은 공개 근거와 맞고, 'KT는 전부 규칙만 쓴다'는 강한 명제는 성립하지 않는다.

## 5b.2 질문②「진짜 Agentic 레퍼런스가 있나?」

**질문 원문:** 진짜 Agentic AI(말도 하고, 툴도 호출하고, 판단하고) 로 레퍼런스를 잡은 업체가 있나?

**정의 [조작적 기준, 국제 단일 표준 아님]:** Agentic = **말**(대화·답변 생성) + **툴 호출**(시스템 조회·실행) + **판단**(다음 행동 선택, 결과를 보고 재계획). 이 세 가지가 **같은 실명 고객·같은 제품·운영 시점**에서 연결될 때만 운영 레퍼런스로 본다. 다른 고객의 '말'·'툴'·'판단'을 합쳐 하나의 레퍼런스로 만들지 않는다.

**판정:** 국내에는 말·툴·판단을 모두 하는 Agentic 운영 레퍼런스가 공개자료로 확인되지 않는다(최근접 = 브리지텍 AoE, 제품 발표 수준). 해외 최근접은 NiCE/Cognigy의 Fabletics 사례(벤더 발표, 실행 로그·재계획 미확인).[68][195]

| 벤더·제품 | 말 | 툴 | 판단·재계획 | 운영 레퍼런스 | 판정 |
|---|---|---|---|---|---|
| **브리지텍 AoE** | ○ 주장 | ○ 주장 | △ 주장 | 우리은행 AX Agent 구축(AoE 여부 미표기).[197] | **데모·발표뿐 — 국내 최근접**.[68] |
| KT 상담Assist | ○ 확인 | × | × | 자사 100번 센터. 「Agentic AICC」는 2026-05 실적발표의 고도화 계획.[192][194] | 일부만(말) |
| LG CNS RT-Advisor | ○ | × | × | a:xink(LG디스플레이)는 사무 플랫폼 — 상담 아님.[198] | 일부만(말).[57] |
| 페르소나 SONA | ○ | × | × | 실명 고객 미연결 | 일부만(말).[85] |
| SKT·삼성SDS·카카오·네이버·마음AI | △ | × | × | 상담지원·자동응대 범위.[61][62][56][60][81][65] | 일부만 |
| NiCE/Cognigy — Fabletics(해외) | ○ | ○ 인증·주문 수정·환불 | △ 행동 순서 선택 서술 | 실명, 가동일 미표기. 벤더 발표 | **해외 최근접 운영 후보**.[195] |
| Genesys AVA(해외) | ○ | △ | △ | Riachuelo·Zeppelin(2026-09-02 발표) | 운영 사례, Agentic 세부 미확인.[196] |

표기: ○ 공개 설명(벤더 주장 포함), △ 일부·미검증, × 증거 미확보. '없음'은 공개자료에서 확보하지 못했다는 뜻이며 기능 부재나 시장 전체 미도입의 증명이 아니다.

- **Fabletics의 한계:** 대화·고객 인증·주문 조회/수정·환불과 의도에 맞는 행동 순서 선택이 같은 사례에서 설명되지만, 고객별 LLM 생성 호출 기록·권한/승인·실패 후 재계획·고객 자체 원출처는 미확인. 벤더 사례만으로 독립 확정하지 않는다.[195]
- **[의견] 경쟁 포지셔닝:** '우리만 진짜 Agentic' 주장은 보류. 국내 상담사 지원과 해외 고객 자율응대는 시장 범위가 다르며, 차별화는 보안·개인정보·기존 화면·업무 승인·예외 복구 성능으로 검증.
- **PoC 합격조건 [제안]: 같은 시나리오의 실행 기록으로 재계획을 입증** — 생성 호출·툴 실행·결과 확인·다음 행동 변경(또는 안전한 보류·이관)이 한 기록에 연결될 때만 'Agentic'으로 인정. 도구 목록(MCP) 등록·정상 경로 시연·단순 재시도만으로는 인정하지 않는다.

# 6. 대표님 가설 4개 — 지지·반대 증거·판정 문턱

## 6.1 가설① 원문 범위와 명제별 판정

**원문 취지:** 프라이빗 Vertex AI/Azure OpenAI가 있어도 개인정보 우려 때문에 현장에 적용하지 못하는 수요가 있는가?

- **보안 수요 명제:** 데이터별 법률·심의·감사 조건은 확인. 보안이 실제 구매 병목의 어느 정도인지는 미측정.[8][7]
- **일부 고객 미도입 명제:** 두 플랫폼을 검토한 특정 고객의 실제 반려 기록·설정·예산이 없어 **미확인**. '현장 AI 전면 미도입'이라는 더 강한 명제로 바꿔 반증하지 않음.
- **전칭 미도입 명제에 대한 반대 증거:** KT·케이뱅크 등의 믿:음 기반 실시간 추천·요약 적용은 공개. 공급자 발표 전달 기사이며 독립 성능 재현 아님. Vertex/Azure 또는 원문 PII 반출 사용 증거도 아님.[87]
- **NH 사례의 범위:** 2026-08-10 NH농협은행 AICC 구축·AI 콜봇·업무 연동은 확인. 기사 속 생성형 상담 지원 강화는 후속 계획 표현이므로 동일 기능 운영 완료로 확대하지 않음.[82]
- **지지 증거의 한계:** 금융위 2026-04-15 지정 169건·출시 45개와 영향도별 절차는 승인 부담 및 실제 사용 경로가 공존함을 보여 줌. **그 중 상담 용도 수는 미확인**, 수요 비중·미도입률·보급률로 환산하지 않음.[7]
- **사업 판정:** 온프렘·보안 우선의 계정별 수요는 조건부 지지, 프라이빗 클라우드 대비 독립 모듈 구매수요는 고객 실사 전 미확인. 기존 SKT[62]·현대커머셜[81] 사례는 경쟁 운영 참고에만 유지하고 가설① 직접 반증에서 제외.

**채택·기각 문턱 [제안, 실측·통계적 시장 추정 아님]**

1. 서로 다른 고객사 10곳의 보안·운영·구매 책임자 인터뷰에서 6곳 이상이 PII/승인/망 문제를 상위 3개 병목으로 제시하고, 적법한 입력 경로를 가진 2곳 및 예산이 있는 유료 의향 2곳을 확보하면 제한 PoC 진행.
2. 승인된 Vertex/Azure 또는 기존 공급자가 동일 통제·연동·전체 비용을 충족하고 추가 모듈 예산 의향이 없으면 독립 온프렘 제품 가설 보류. 보안 문제 자체가 없어졌다는 뜻은 아님.
3. 서면 데이터/계약 권한을 얻지 못하거나 총 기여이익이 음수면 진행 중단. 고객사 1곳의 여러 BPO를 독립 수요 10곳으로 중복 집계 금지.

## 6.2 가설②~④

| 가설 | 판정 | 지지 증거 | 한계·반대 조건 | 검증·중단 기준 [제안] |
|---|---|---|---|---|
| ② 사내 PII 금고·토큰 호출·사내 복원 | 기술적 수단으로 유효, 완결 해법·규제 면제는 아님 | 복원 가능한 mapping/encrypt 설계.[21][35] | 맥락의 간접 식별, 로그·검색, 타 세션 복원 및 법적 의무 잔존.[4] | 한국어 PII·부분 전사 수정·동시 복원·유효기간 시험. 중대 원문 반출/교차 복원 1건이면 반출 중단; 표본 0건은 무위험 증명 아님 |

| 가설 | 판정 | 지지 증거 | 한계·반대 조건 | 검증·중단 기준 [제안] |
|---|---|---|---|---|
| ③ Jev-like 경량 판단으로 필요한 때만 LLM | 설계 조건부 지지, 오픈소스 Jev 전제 미성립 | 로컬 의미 분기·모델 라우팅 코드 존재.[18][19][20] | Jev SDK≠가중치, 기본 계약 hosted API. 한국어·교정된 확률·총원가 미검증.[45][30] | 중요 개입 재현율 99% 목표+신뢰 구간, 품질 비열등성, 전체 비용 검증. 판단 보류는 사람·규칙·상위 모델로 승격 |
| ④ 기존 STT·PC 채널·NEVA 재사용 | 선택 도입 지지,2소스=2역할/화면 무누락 가정은 조건부 | NEVA 다중 앱 맥락 취득 공식 설명.[48] | loopback 재생 혼합, SIPREC 혼합 구성 가능. 화면 전사 API/SLA 미확인.[28][42] | 기존 STT API 우선. 화면 일치≥99.9%·이벤트 p95≤500ms·역할 혼동≤1% 목표. 타 통화 혼합 1건 중단, 불확실시 차단·알림 |

출처: 행별 공식 문서·법령. 문턱은 연구자의 실험 설계이며 고객 위험 허용도·유형별 표본 수·사전 등록 기준에 따라 승인 후 조정.

### 6.3 중요한 비대칭 [의견]

- 형식이 맞아도 판단이 맞는 것은 아님. 출력 logprob·유사도·Jev 확률은 한국어에서 교정된 정답 확률과 다름.[32]
- 게이트 오탐은 비용·알림 피로, 미탐은 중요 개입 누락. 전체 정확도 하나로 상쇄하지 않음.
- PII는 평균 마스킹 정확도 대신 주민번호·계좌·건강·간접 식별 및 세션별 누출을 따로 검사.

## 7. 진입 전략·가격·파트너 [의견·가정]

### 7.1 우선순위

1. 기존 STT·업무 접근권·예산을 가진 민간 고객사/BPO의 읽기·권고 지원부터. 대표님의 관계는 접근성 가설이지 판매 확약이 아님.
2. 금융·보험은 지정/예외·데이터·보안 조건이 확인된 고객 및 SI 파트너 경로. 실명 도입 사례가 있다는 것과 새 벤더 승인 가능성은 다름.
3. 공공은 초기 SOM에서 보수적으로 제외, 공동 기반·인증·조달 파트너 경로 검토.

### 7.2 가격 비교 기준

| 비교 항목 | 확인 가격·조건 | 본 제품과 비교할 때 주의 |
|---|---|---|
| 해외 상담사 Copilot 추가 기능 | Zendesk Copilot USD50/상담사/월, 연간 결제. 확인일 2026-09-30 공식 한국어 페이지.[92] | 기반 구독·사용량·음성·연동 별도. 온프렘 음성 Copilot 동일 범위 아님 |

| 비교 항목 | 확인 가격·조건 | 본 제품과 비교할 때 주의 |
|---|---|---|
| 해외 컨택센터 전체 플랫폼 | Genesys CX USD75~240, NiCE 패키지 USD110~249/사용자·월.[73][76] | CCaaS 전체 패키지, 토큰/세션·계약 조건 별도. Copilot 단품 가격으로 사용 금지 |
| 국내 RT-Advisor·Assist·SONA | 공개 단품 가격 **미확인**.[57][59][85] | 동일 좌석/사용량/구축·보안 범위의 서면 견적 필요 |
| NH농협은행 AICC 사업 | 약 400억원 전체 구축사업,2026-08-10 발표.[82] | 등록 좌석·모듈·기간 불명, 단품 월단가로 나눌 수 없음 |
| TheNoun 견적 실험 | 월 3/6/10만원/등록 사용자, **연구자 가정** | 검증된 시장 가격 아님. 호출·STT·GPU·구축·보안·유지보수 포함 범위 명시 |

**비교 절차 [제안]:** 고객사 100등록 좌석·동시 사용 비율·통화량을 먼저 확인 → 기능·배포·SLA·계약 기간·VAT·초과 과금·어댑터/보안 인력을 맞춘 3년 전체 비용 비교. 원화/달러 단순 우열 계산 금지; 비교일 환율 및 실제 견적 확보 후 계산. 국내 단가 미확인 칸을 해외 패키지 요금으로 채우지 않음.

- 기여이익 = SW·사용량 수익 - 모델/STT/GPU/통신 - 고객별 운영·지원 변동비. 구축비 회수·법무/심의·영업비 별도.
- ROI 예시 **전부 가정**:100좌석×60통화/일×20일×30초÷3600=1000시간. 시간당 1.8만원=1800만원 잠재 가치,40% 현금화=720만원. SW600만원 차감 120만원은 STT·인프라·운영 비용 포함 전. 절감 시간≠현금 이익 또는 즉시 감원 가능성. wiki/v2-calculations

## 7.3 파트너와 역할

| 상대 | 기대 역할 [제안] | 서면 확인할 항목 |
|---|---|---|
| 고객사+BPO | 데이터 책임·지식·예산·상담 운영·효과 측정 | 비용 부담/효익 수취, 상담사 안내, 학습·외부 반출 권한 |
| SI/IPCC·보안 | 설치·망·감사·장애·심의 | API·재위탁·배포·운영 책임 및 제휴 권리 |
| STT 공급자 | 통화ID·채널·부분/최종·수정 이벤트 | 실시간 API·한국어 품질·라이선스·SLA |
| NICE/알피에이 | NEVA 어댑터·지원·유지보수 | 버전·커넥터·VDI/OCR·개발/연동 권리·외부 전송 범위 |
| 모델 공급자 | 가중치·상업 공급권·지원 | 내부 사용 vs 고객 설치/호스팅 판매, 보존·구성 목록 |

접촉·계약은 미실행.

### 7.4 현대차 현장 PoC

현대차 현장은 PoC 현장으로 확보 가능한 후보이며, 확보한 실증 고객은 아님. 아래는 현장 활용 시 확인할 범위.

| 현장 활용 후보 [제안] | 확인할 문서·권한 | 승인 전 범위 |
|---|---|---|
| 가설①의 보안/망/PII 병목 인터뷰 | 현대차/원·수탁사의 정당한 인터뷰·기밀 공개 범위 | 업무 경험을 판매 확약으로 사용 금지, 내부 보안 자료 반출 금지 |
| 가설④ STT·NEVA 읽기 전용 시험 | 고객사 승인권자, STT/NEVA 계약·연동/실데이터·PC 설치/화면 권리 | 합성 통화·합성 화면·독립 시험환경만. 알피에이 승인만으로 현대차 권한 추정 금지 |
| 코드·프롬프트·RAG·가이드 재사용 | 배경/성과 IP·제3자 권리·양도/이용허락·비밀 관리 | 작성자 개인 소유 또는 대가 지급으로 모든 권리 이전이라는 자동 결론 금지 |

**법문 근거:** 영업비밀의 비공지성·경제가치·비밀 관리 및 계약상 의무, 저작권법의 업무상저작물·프로그램 공표 예외·양도/허락 범위는 실제 계약과 대조할 것.[153][154]

개인정보 재위탁이면 위탁자 동의 등 법제26조를 별도 검토.[149]

**진입 선택:** 현대차는 승인된 테스트베드 가능성을 확인할 후보이지 확보한 실증 고객이 아님. 계약/기밀상 불가능하면 독립 민간 계정·합성환경으로 검증. 고객 명칭을 외부 레퍼런스로 쓰는 것도 별도 승인 전제. 실제 계약은 담당 법무 서면 검토 필요.

## 8. 리스크·결론이 바뀌는 조건

| 위험 | 조기 지표 | 대응·사업 보류 조건 [제안] |
|---|---|---|
| 경쟁 동질화 | 기존 벤더가 동일 통제·연동·낮은 전체 비용 제공 | 독립 제품 보류 또는 통합 파트너 전환 |
| 데이터·연동 권리 | API·STT·NEVA·고객사 승인 부재 | 후보 변경, 승인 없는 화면·녹취 수집 금지 |
| PII·교차 복원 | 원문 반출·로그 노출·다른 통화 토큰 복원 | 즉시 차단·사고 대응·재시험 |
| 게이트 품질 | 중요 미탐·오탐·판단 보류 급증 | 고위험은 규칙 강제 승격, 판단 범위 축소 |
| 환각·위험 행동 | 근거 없는 추천·권한 밖 실행 | 권고 모드·근거·결정론적 권한/승인·중지 |
| 경제성 | GPU 유휴·전사/연동비가 호출 절감 초과 | 가격/범위 재설계; 음의 기여이익이면 확장 보류 |
| 공급권 | NC·경쟁 서비스·제3자 공급 조건 미해소 | 서면 확인 전 유료 PoC·재판매 금지 |
| 화면 변경 | 앱·스크롤·VDI 변경 때 누락·중복 | API 전환·지원 범위 고정·불확실시 차단 |
| 계약·노사 | IP·모니터링 범위 미확인 | 계약/법무/노사 확인 전 실데이터 단계 보류 |
| 심의·시장 | 승인 대기·예산 부재·단일 TAM 과신 | 고객별 bottom-up SAM, 확정 계약만 SOM 반영 |

연구자 위험 분석이며 특정 벤더 사고·결함을 주장한 표가 아님.

# 9. 다음90일 실행안 — 승인 대기·인력·비용 포함 [제안]

## 9.1 일정과 승인 병목

- **90일은 기술/구매 검증의 상한 목표이지 금융 지정·공공 조달·보안 심의 완료 약속이 아님.** 개별 법정 처리 기간과 실제 고객 심의 일정은 미확인.
- 고객사 법무·정보보호·노사·조달 검토에 **2~6주 병렬 버퍼를 가정**. 산업 평균이나 법정 기한 아님. 신규 지정·추가 심의가 필요하거나 승인 지연 시 종료일을 재계획.
- 승인이 없으면 합성 데이터·오프라인 재현까지만. 46일이 되었다는 이유만으로 실통화 shadow/live를 시작하지 않음.

| 기간 | 실행 | 승인·산출물·통과 문턱 |
|---|---|---|
| 1~15일 | 서로 다른 고객사 10곳 인터뷰,2곳 후보, 계약·API·예산 책임자 확인 | 가설① 병목/예산 실사, 데이터 지도·역할 표. 현대차/알피에이 현장 사용은 별도 서면 승인 전제 |
| 16~30일 | 합성/적법한 비식별 상담 3000개, PII/게이트 실험 설계; 법무·보안 심의 병렬 준비 | 통화 단위 학습·교정·시험 분리, 중요 유형별 표본 수·신뢰 구간. 중대 누출 1건이면 반출 경로 보류 |
| 31~45일 | 허가된 STT API1종+권고 검색 1종, NEVA 읽기 전용 별도 기술 시험 | 화면 누락≤0.1%·일치≥99.9%·이벤트 p95≤500ms 목표. 타 통화 혼합 1건 중단. 미승인시 합성 화면만 |
| 46~60일 | 승인 완료 계정만 상담사 20~50명 shadow; 전량 호출·규칙·로컬 모델 비교 | 중요 개입 재현율≥99% 목표+신뢰 구간, 품질 비열등성·전체 비용. 미달은 규칙 승격/범위 축소 |
| 61~75일 | 고객사·상담사 운영 승인 후 제한 live/교차 시험 | AHT 또는 ACW≥10% 개선 목표, 품질·알림 피로 악화 시 중단. 자동 거래 권한 제외 |
| 76~90일 | 2곳 유료 전환 후보·독립 견적·지원 원가 확정 | 유료 의향 0·기여이익 음수·권리 미확인 시 확장 보류. 심의 미완료는 기술 성공과 구분 |

모든 좌석·표본·목표는 실적·산업 표준이 아닌 사전 승인할 실험 가정. 개인정보 유형별 0건이 무위험 보장은 아님.

**법정 절차와 계획 버퍼 구분:** AI기본법 시행령제25조 고영향 확인요청은 30일 회신·한 차례 최대 30일 연장. 개인정보법 시행령제44조의3 자동화 결정 거부/설명 요구 처리 기간도 별도 권리행사 절차. **두 기간 모두 고객 보안심의 SLA가 아님.** 고객 실제 일정·금융/공공의 해당 신청은 별도 확인.[147][150]

## 9.2 인력·예산 시나리오 [가정·견적 미확인]

| 자원 | 가정 투입 | 책임 |
|---|---|---|
| PM/도메인 | 3인주 | 계정·계약·상담 정책·승인/일정 |
| 백엔드/연동 | 8인주 | STT·금고·NEVA·감사·장애 |

| 자원 | 가정 투입 | 책임 |
|---|---|---|
| ML/데이터 | 6인주 | 라벨·한국어 PII·게이트·교정·평가 |
| 보안/법무 대응 | 3인주 | 흐름·반출·권리·심의 문서 |
| QA/상담 평가 | 4인주 | 유형별 위험·성능·사용성·회귀 |

- 총 24인주, 실제 고정 인원은 역할 겸임과 가용 시간 확인 후 결정. 약 2~3명 핵심 투입+파트타임 전문가 구성 **제안**, 현재 채용/가용 인력 미확인.
- 완전원가 1인주 200~350만원 **가정** → 인력 4800~8400만원. 대표 노동을 무료로 계산하지 않은 기회비용 포함; 실제 현금 지출과 다름.
- GPU/서버 3개월 300~900만원, 외부 법무/보안 300~1000만원, STT/모델/시험 100~300만원 **가정**. GPU 구매가·클라우드 정가·공급자 견적 아님. 모델·양자화·동시성 실측 후 장비/임대 선택.
- 합계 5500~1억 600만원, 예비비 15% 포함 **6325~1억 2190만원**. 세금·하드웨어 구매·보증/장기 운영 포함 범위는 견적 후 재산정. 실제 예산 상한 승인 없으면 고객 인터뷰·합성 검증으로 축소. wiki/v2-calculations

## 10. 미확인·후속 정보 요청

1. **시장:** 최신 전국 상담사·좌석·업종, 좁은 Copilot 매출·성장·구매 의향·국내 단품 가격. TAM 민감도는 유료 도달 시장의 확정값 아님.
2. **가설①:** 실제 Vertex/Azure 검토·반려 기록, PII/승인 병목 순위, 상담 용도 금융 지정/출시 수, 해당 모듈 예산.
3. **규제:** 개별 혁신금융 지정 조건·시행세칙 별표7 전체·CSAP/기관 보안·AI기본법 해당성·고지·상담사 모니터링·노사 검토.
4. **PII:** 한국어 이름/주소·STT 변형·금고 동시성/복원·국외 접근/보존·해당 처리 법무 의견.
5. **게이트/모델:** 한국어 현장 성능·교정·모델/코드/학습 데이터 권리·GPU·지연·가중치의 고객 상업 공급권.
6. **STT/NEVA:** 리턴제로/네이버 등 정확한 계약 상품, 온프렘 옵션·실시간 API·음성 경로·역할 분리·버전/커넥터·VDI/OCR·지원 SLA.
7. **현대차 현장:** 고객사 계약의 비밀·성과물 IP·실데이터/화면/NEVA 사용권·재위탁·고객 명칭 공개 권한.
8. **경쟁/접근:** 다수 최신 모델·SKU·가격·가역 금고 미공개. NICE/Verint 국내 실명 Copilot·과거 사례의 현행 운영 미확인. 실패/로그인 필요 자료는 검색문으로 확정 대체하지 않음.
9. **일정/예산:** 심의 실제 대기·법정 기한·고객 보안 요구·인력·장비/자문 견적. 90일 live 및 유료 전환 보장 아님.

## 11. 조사 증빙·재검수 인계

- v1은 보존, v2의 출처·원문 단서·계산은 볼트의 Markdown에서 확인하도록 구성. 로컬 JSON/임시 캐시가 없다는 이유로 검수가 막히지 않도록 본문 및 wiki/v2-calculations에 산식 제공.
- 원출처별 raw → 개념 wiki → index/log → report 구조. 공개 자료 본문 확인과 공급자 주장, 계산 가정, 실험 제안 구분.
- 모델 실행·한국어 품질 측정·현장 PoC·고객 접촉·실제 계약 검토·신규 승인은 수행하지 않음. 라이선스 검토는 확인한 ID/버전 범위.
- 인용 번호·파일·링크·산술 검증 결과는 v2-verification에 기록. **독립 재검수·대표 보고 승인 담당은 하린**, 작성자가 스스로 통과 판정하지 않음.

관련: wiki/market-sizing · wiki/vendors-draft · wiki/technology-draft · wiki/regulation-draft · wiki/v2-cloud-pricing-hypothesis · wiki/v2-calculations · wiki/v2-model-licenses-integrated · wiki/v2-gating-stt-integrated · wiki/v2-regulation-employment-integrated · wiki/v2-source-catalog

## Sources


[1] https://kosis.kr/civilComplaint/qnaDetail.do?boardIdx=21682 — market

[2] http://kctuli.kctu.org/bbs/read.php?tbl=work&seq=551 — market

[3] https://www.newsprime.co.kr/news/article/?no=611530 — market

[4] https://www.law.go.kr/LSW/lsInfoR.do?lsiSeq=283839&chrClsCd=010202&efYd=20260911 — 개인정보 보호법 시행 2026-09-11

[5] https://www.law.go.kr/LSW/lsInfoR.do?lsiSeq=283841&chrClsCd=010202&efYd=20260911 — 신용정보법 시행 2026-09-11

[6] https://www.law.go.kr/LSW/admRulLsInfoR.do?admRulSeq=2100000282622&admRulId=21828&joTpYn=Y&languageType=KO&chrClsCd=010202 — 전자금융감독규정 시행 2026-07-15

[7] https://www.fsc.go.kr/no010101/86712 — 생성형 AI 혁신금융서비스 모델변경 절차 간소화

[8] https://www.fsc.go.kr/no010101/86745 — 내부 업무망 SaaS 망분리 예외 시행

[9] https://www.fsc.go.kr/no010101/87646 — 제2차 망분리 규제 긴급완화: 보안목적 테스트

[11] https://isms-p.or.kr/ntcn/rcsrm/selectGnrlVrtlRcsrmDetail.do?searchRcsrmMngId=RCSRMID_000000010181 — CSAP FAQ 2026.07 (PDF 본문 확인)

[12] https://isms-p.or.kr/ntcn/rcsrm/selectGnrlVrtlRcsrmDetail.do?searchRcsrmMngId=RCSRMID_000000010180 — CSAP 보안인증기준 해설서 2026.07

[14] https://www.mois.go.kr/frt/bbs/type010/commonSelectBoardArticle.do?bbsId=BBSMSTR_000000000008&nttId=126757 — 공공부문 AI 도입·활용 가이드 배포

[15] https://www.law.go.kr/LSW/admRulLsInfoR.do?admRulSeq=2200000108629&joTpYn=Y&languageType=KO&chrClsCd=010202 — 전자금융감독규정시행세칙 시행 2026-04-20

[16] https://pipc.go.kr/np/cmm/fms/FileDown.do?atchFileId=FILE_000000000559959&fileSn=3 — 생성형 인공지능 개발·활용 개인정보 처리 안내서 PDF 2025.08

[18] https://raw.githubusercontent.com/aurelio-labs/semantic-router/main/README.md — semantic-router

[19] https://raw.githubusercontent.com/lm-sys/RouteLLM/main/README.md — route-llm

[20] https://raw.githubusercontent.com/vllm-project/semantic-router/main/README.md — vllm-router

[21] https://raw.githubusercontent.com/microsoft/presidio/main/docs/samples/python/pseudonymization.ipynb — presidio-pseudo

[22] https://raw.githubusercontent.com/microsoft/presidio/main/docs/analyzer/languages.md — presidio-languages

[23] https://raw.githubusercontent.com/microsoft/presidio/main/presidio-analyzer/presidio_analyzer/conf/default_recognizers.yaml — presidio-recognizers

[25] https://huggingface.co/LGAI-EXAONE/EXAONE-4.0-32B/raw/main/LICENSE — exaone-license

[26] https://huggingface.co/LGAI-EXAONE/K-EXAONE-236B-A23B/raw/main/LICENSE — kexaone-license

[27] https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0/raw/main/README.md — upstage

[28] https://learn.microsoft.com/en-us/windows/win32/coreaudio/loopback-recording — loopback

[29] https://docs.typesafe.ai/introduction.md — jev-intro

[30] https://docs.typesafe.ai/models.md — jev-models

[32] https://docs.typesafe.ai/model-jaggedness/jev-1.13.md — jev-jagged

[34] https://raw.githubusercontent.com/data-privacy-stack/presidio/main/README.MD — presidio-readme

[35] https://raw.githubusercontent.com/data-privacy-stack/presidio/main/docs/anonymizer/index.md — presidio-anon

[36] https://raw.githubusercontent.com/data-privacy-stack/presidio/main/LICENSE — presidio-license

[37] https://raw.githubusercontent.com/aurelio-labs/semantic-router/main/LICENSE — semantic-license

[38] https://raw.githubusercontent.com/lm-sys/RouteLLM/main/LICENSE — route-license

[39] https://raw.githubusercontent.com/vllm-project/semantic-router/main/LICENSE — vllm-license

[40] https://www.nice.com/faq/products-faqs/what-is-robotic-process-automation-neva-and-how-does-it-streamline-agent-tasks — nice-faq

[41] https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-recognize-speech-multichannel — multichannel

[42] https://www.rfc-editor.org/rfc/rfc7866.txt — siprec

[44] https://huggingface.co/LGAI-EXAONE/K-EXAONE-236B-A23B/raw/main/README.md — kexaone-card

[45] https://typesafe.ai/legal/mca — jev-mca

[47] https://arxiv.org/html/2406.18665v4 — route-paper4

[48] https://resources.nice.com/wp-content/uploads/2024/03/rpa-nice-solution-e-book.pdf — nice-studio

[50] https://huggingface.co/upstage/SOLAR-10.7B-v1.0/raw/main/README.md — solar-base

[51] https://raw.githubusercontent.com/data-privacy-stack/presidio/main/docs/project_transition.md — presidio-transition

[52] https://learn.microsoft.com/en-us/samples/microsoft/windows-classic-samples/applicationloopbackaudio-sample — process-loopback

[53] https://vllm-sr.ai/docs/intro — vllm-intro

[54] https://raw.githubusercontent.com/typesafe-ai/system-one-adapter-python/main/LICENSE — adapter-license

[55] https://raw.githubusercontent.com/data-privacy-stack/presidio/main/presidio-analyzer/presidio_analyzer/predefined_recognizers/country_specific/korea/kr_rrn_recognizer.py — kr-rrn

[56] https://www.samsungsds.com/kr/news/aicc_210509.html — samsung

[57] https://connect.lgcns.com/language-masters/ko/service/ai/ai-contact-center — lgcns

[58] https://www.lgcns.com/kr/newsroom/press/detail.ko_0809 — lgcns-cases (시점 미상)

[59] https://gov.kt-aicc.com/user/assist — kt

[60] https://kakaoenterprise.com/press/kakaoiconectcentersaas — kakao

[61] https://news.sktelecom.com/208852 — skt

[62] https://news.sktelecom.com/230764 — skt-live

[63] https://www.skelterlabs.com/blog/bella-qna-assistantchatbot — skelter

[64] https://www.skelterlabs.com/usecase — skelter-cases

[65] https://fast-aicc.maum.ai/intro?lang=ko — maum

[66] https://www.personaai.co.kr/news/view/93 — persona

[67] https://www.personaai.co.kr/company/about — persona-cases

[68] https://www.bridgetec.co.kr/v2/solution/aoe — bridge

[69] https://www.bridgetec.co.kr/investment/ir/index.jsp — bridge-cases

[70] https://www.nexus.co.kr — nexus

[71] https://saltlux.com/kor/contents/view.do?mId=34 — saltlux

[72] https://www.genesys.com/ko-kr/blog/post/genesys-cloud-agent-copilot-deep-dive — genesys

[73] https://www.genesys.com/ko-kr/pricing — genesys-price

[74] https://help.genesys.cloud/faqs/category/artificial-intelligence — genesys-ai

[75] https://www.genesys.com/company/newsroom/announcements/genesys-cloud-reaches-nearly-1-3-billion-annual-recurring-revenue-in-third-quarter-of-fiscal-year-2024 — genesys-cases

[76] https://www.nice.com/pricing — nice-price

[77] https://help.nicecxone.com/content/admin/analyticsservices/analyticspolicies.htm?TocPath=Admin%7CAdmin%7CAnalyticsServices%7C_____3 — nice-pii

[78] https://www.nice.com/generative-ai-in-customer-service/generative-ai-governance — nice-governance

[79] https://www.verint.com/agent-virtual-assistant/?vrntabp=calabrio — verint

[80] https://www.verint.com/wp-content/uploads/2024/10/verint-supported-commercial-LLMs.pdf — verint-models

[81] https://newsroom.hyundaicard.com/front/board/%ED%98%84%EB%8C%80%EC%BB%A4%EB%A8%B8%EC%85%9C,-%EB%84%A4%EC%9D%B4%EB%B2%84%ED%81%B4%EB%9D%BC%EC%9A%B0%EB%93%9C-%EA%B8%B0%EB%B0%98-AI-%EC%83%81%EB%8B%B4%EC%9B%90-%EC%84%9C%EB%B9%84%EC%8A%A4-%EA%B5%AC%EC%B6%95 — naver-case

[82] https://www.newspim.com/news/view/20260810000291 — kt-latest

[83] https://www.nice.com/products/copilot-for-agents — nice-copilot

[84] https://www.verint.com/embrace-generative-ai — verint-llm

[85] https://www.personaai.co.kr/products/sona_assistant — persona-assist

[86] https://www.personaai.co.kr/products/kgpt — persona-kgpt

[87] https://www.asiae.co.kr/article/2025073009183804420 — kt-model

[88] https://cloud.google.com/vertex-ai/generative-ai/docs/data-governance — Gemini Enterprise Agent Platform and zero data retention  |  Google Cloud Documentation

[89] https://cloud.google.com/vertex-ai/docs/general/vpc-service-controls — VPC Service Controls with Gemini Enterprise Agent Platform  |  Google Cloud Documentation

[90] https://learn.microsoft.com/en-us/azure/foundry/responsible-ai/openai/data-privacy — Data, privacy, and security for Foundry Models sold by Azure in Microsoft Foundry - Microsoft Foundry | Microsoft Learn

[91] https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/network — Securing Azure OpenAI inside a virtual network with private endpoints (classic) - Microsoft Foundry (classic) portal | Microsoft Learn

[92] https://www.zendesk.com/pricing — Zendesk 가격 플랜 | 월 $19부터

[93] https://huggingface.co/K-intelligence/Midm-2.0-Base-Instruct/resolve/35479c5fc9a18a5db7cc6dbadcf1db68db7beab0/LICENSE.txt — K-intelligence/Midm-2.0-Base-Instruct LICENSE.txt

[94] https://huggingface.co/K-intelligence/Midm-2.0-Base-Instruct/resolve/35479c5fc9a18a5db7cc6dbadcf1db68db7beab0/README.md — K-intelligence/Midm-2.0-Base-Instruct README.md

[95] https://huggingface.co/K-intelligence/Midm-2.0-Mini-Instruct/resolve/383eb221c52a32278f1985257b264ade8d982e60/LICENSE.txt — K-intelligence/Midm-2.0-Mini-Instruct LICENSE.txt

[96] https://huggingface.co/K-intelligence/Midm-2.0-Mini-Instruct/resolve/383eb221c52a32278f1985257b264ade8d982e60/README.md — K-intelligence/Midm-2.0-Mini-Instruct README.md

[97] https://huggingface.co/skt/A.X-4.0/resolve/a3393fd67bf09a34a85e66dfd2c2da61c4497b67/LICENSE — skt/A.X-4.0 LICENSE

[98] https://huggingface.co/skt/A.X-4.0/resolve/a3393fd67bf09a34a85e66dfd2c2da61c4497b67/README.md — skt/A.X-4.0 README.md

[99] https://huggingface.co/skt/A.X-4.0-Light/resolve/ba21c20ea1b31ded1ec3e2fb432335077dc4be98/LICENSE — skt/A.X-4.0-Light LICENSE

[100] https://huggingface.co/skt/A.X-4.0-Light/resolve/ba21c20ea1b31ded1ec3e2fb432335077dc4be98/README.md — skt/A.X-4.0-Light README.md

[101] https://huggingface.co/kakaocorp/kanana-1.5-15.7b-a3b-instruct/resolve/df9d415c38e198149a65957cb4dec6cf2a7ba162/LICENSE — kakaocorp/kanana-1.5-15.7b-a3b-instruct LICENSE

[102] https://huggingface.co/kakaocorp/kanana-1.5-15.7b-a3b-instruct/resolve/df9d415c38e198149a65957cb4dec6cf2a7ba162/README.md — kakaocorp/kanana-1.5-15.7b-a3b-instruct README.md

[103] https://huggingface.co/Qwen/Qwen3-8B/resolve/b968826d9c46dd6066d109eabc6255188de91218/LICENSE — Qwen/Qwen3-8B LICENSE

[104] https://huggingface.co/Qwen/Qwen3-8B/resolve/b968826d9c46dd6066d109eabc6255188de91218/README.md — Qwen/Qwen3-8B README.md

[105] https://huggingface.co/google/gemma-3-12b-it/resolve/96b6f1eccf38110c56df3a15bffe176da04bfd80/README.md — google/gemma-3-12b-it README.md

[106] https://huggingface.co/naver-hyperclovax/HyperCLOVAX-SEED-Think-14B/resolve/9b74e35d4c7e4ffec489f4171273caca8948a2b9/LICENSE — naver-hyperclovax/HyperCLOVAX-SEED-Think-14B LICENSE

[107] https://huggingface.co/naver-hyperclovax/HyperCLOVAX-SEED-Think-14B/resolve/9b74e35d4c7e4ffec489f4171273caca8948a2b9/README.md — naver-hyperclovax/HyperCLOVAX-SEED-Think-14B README.md

[108] https://huggingface.co/kakaocorp/kanana-1.5-8b-instruct-2505/resolve/c963a5f4f6496c749f94064a20b33028b0db9f19/LICENSE — kakaocorp/kanana-1.5-8b-instruct-2505 LICENSE

[109] https://huggingface.co/kakaocorp/kanana-1.5-8b-instruct-2505/resolve/c963a5f4f6496c749f94064a20b33028b0db9f19/README.md — kakaocorp/kanana-1.5-8b-instruct-2505 README.md

[110] https://huggingface.co/kakaocorp/kanana-1.5-2.1b-instruct-2505/resolve/7df4bc35ccd610e451809d7106e1c3cf82bfd44c/LICENSE — kakaocorp/kanana-1.5-2.1b-instruct-2505 LICENSE

[111] https://huggingface.co/kakaocorp/kanana-1.5-2.1b-instruct-2505/resolve/7df4bc35ccd610e451809d7106e1c3cf82bfd44c/README.md — kakaocorp/kanana-1.5-2.1b-instruct-2505 README.md

[112] https://ai.google.dev/gemma/terms — https://ai.google.dev/gemma/terms terms

[113] https://ai.google.dev/gemma/prohibited_use_policy — https://ai.google.dev/gemma/prohibited_use_policy prohibited_use_policy

[114] https://docs.vllm.ai/en/v0.15.0/features/structured_outputs — vllm-structured

[115] https://docs.vllm.ai/en/v0.12.0/api/vllm/logprobs — vllm-logprobs

[116] https://raw.githubusercontent.com/vllm-project/vllm/main/LICENSE — vllm-license

[117] https://raw.githubusercontent.com/vllm-project/vllm/v0.15.0/vllm/sampling_params.py — vllm-sampling-params

[118] https://huggingface.co/answerdotai/ModernBERT-base/raw/main/README.md — modernbert-card

[119] https://huggingface.co/monologg/koelectra-base-v3-discriminator/raw/main/README.md — koelectra-card

[120] https://raw.githubusercontent.com/monologg/KoELECTRA/master/LICENSE — koelectra-license

[121] https://raw.githubusercontent.com/meta-llama/PurpleLlama/main/Llama-Guard3/8B/MODEL_CARD.md — llamaguard-public-card

[122] https://huggingface.co/meta-llama/Llama-Guard-3-8B/blob/main/LICENSE — llamaguard-hf-license-page

[123] https://raw.githubusercontent.com/meta-llama/PurpleLlama/main/Llama-Guard3/8B/LICENSE — llamaguard-public-license

[124] https://raw.githubusercontent.com/NVIDIA/NeMo-Guardrails/main/README.md — nemo-readme

[125] https://raw.githubusercontent.com/NVIDIA/NeMo-Guardrails/main/LICENSE.md — nemo-license-full

[126] https://raw.githubusercontent.com/urchade/GLiNER/main/LICENSE — gliner-code-license

[127] https://huggingface.co/urchade/gliner_multi-v2.1/raw/main/README.md — gliner-multi-card

[128] https://huggingface.co/urchade/gliner_multi_pii-v1/raw/main/README.md — gliner-pii-card

[129] https://huggingface.co/taeminlee/gliner_ko/raw/main/README.md — gliner-ko-card

[130] https://www.rtzr.ai — rtzr-home

[131] https://developers.rtzr.ai/docs/stt-streaming — rtzr-stream-ko

[132] https://developers.rtzr.ai/docs/en/stt-streaming/grpc — rtzr-grpc

[133] https://developers.rtzr.ai/docs/pricing — rtzr-pricing

[134] https://developers.rtzr.ai/docs/stt-file — rtzr-file

[135] https://developers.rtzr.ai/docs/stt-file/diarization — rtzr-diarization-ko

[136] https://developers.rtzr.ai/docs/stt-file/multi-channel — rtzr-multichannel

[137] https://www.rtzr.ai/en/pricing — rtzr-enterprise-pricing

[138] https://api.ncloud-docs.com/docs/ai-application-service-clovaspeech-grpc — clova-grpc

[139] https://api.ncloud-docs.com/docs/ai-application-service-clovaspeech-longsentence — clova-long-api

[140] https://raw.githubusercontent.com/openai/whisper/main/README.md — whisper-readme

[141] https://raw.githubusercontent.com/openai/whisper/main/LICENSE — whisper-license

[142] https://raw.githubusercontent.com/openai/whisper/main/whisper/tokenizer.py — whisper-tokenizer

[143] https://raw.githubusercontent.com/SYSTRAN/faster-whisper/master/README.md — faster-whisper-readme

[144] https://raw.githubusercontent.com/SYSTRAN/faster-whisper/master/LICENSE — faster-whisper-license

[145] https://raw.githubusercontent.com/ufal/whisper_streaming/main/README.md — whisper-streaming

[146] https://www.law.go.kr/법령/인공지능발전과신뢰기반조성등에관한기본법 — 인공지능발전과신뢰기반조성등에관한기본법

[147] https://www.law.go.kr/법령/인공지능발전과신뢰기반조성등에관한기본법시행령 — 인공지능발전과신뢰기반조성등에관한기본법시행령

[148] https://www.law.go.kr/법령/통신비밀보호법 — 통신비밀보호법

[149] https://www.law.go.kr/법령/개인정보보호법 — 개인정보보호법

[150] https://www.law.go.kr/법령/개인정보보호법시행령 — 개인정보보호법시행령

[151] https://www.law.go.kr/법령/근로자참여및협력증진에관한법률 — 근로자참여및협력증진에관한법률

[152] https://www.law.go.kr/법령/근로기준법 — 근로기준법

[153] https://www.law.go.kr/법령/부정경쟁방지및영업비밀보호에관한법률 — 부정경쟁방지및영업비밀보호에관한법률

[154] https://www.law.go.kr/법령/저작권법 — 저작권법

[155] https://law.go.kr/LSW/precInfoP.do?precSeq=215873 — 대법원2006도4981

[156] https://www.law.go.kr/LSW/precInfoP.do?evtNo=2002%EB%8F%84123 — 대법원2002도123

[157] https://raw.githubusercontent.com/NandhaKishorM/laya/6d942c92081fbc139e736bbd9ac0023223c29b7f/README.md — laya-github-readme-pinned

[158] https://raw.githubusercontent.com/NandhaKishorM/laya/6d942c92081fbc139e736bbd9ac0023223c29b7f/LICENSE — laya-github-license-pinned

[159] https://raw.githubusercontent.com/NandhaKishorM/laya/6d942c92081fbc139e736bbd9ac0023223c29b7f/BENCHMARKS.md — laya-benchmarks-pinned

[164] https://huggingface.co/convaiinnovations/laya/raw/55cf4c4ebb4ebe31b2550e8bdf3bd21b99753851/README.md — laya-hf-pinned-card

[165] https://raw.githubusercontent.com/NandhaKishorM/laya/6d942c92081fbc139e736bbd9ac0023223c29b7f/research/results/cpu_51_language_sweep_refreshed.json — laya-refreshed-results-pinned

[179] https://api.github.com/repos/browser-use/jev-ultrafast — GitHub REST API / Browser Use

[180] https://raw.githubusercontent.com/browser-use/jev-ultrafast/main/README.md — Browser Use README

[181] https://vercel.com/blog/ai-gateway-jev-model-launch — Vercel 블로그 2026-09-18

[182] https://vercel.com/changelog/typesafe-ai-jev-now-available-on-ai-gateway — Vercel changelog 2026-09-16

[183] https://journal.qualiteg.com/jev-typesafe-ai-pricing-python-hands-on — Qualiteg 2026-09-29

[184] https://construct.computer/blog/jev-ai-agents — Construct 2026-09-30

[185] https://api.github.com/repos/NandhaKishorM/laya — GitHub REST API / NandhaKishorM

[186] https://www.anth.us/blog/jev-vs-laya — Anthus 2026-09-21

[187] https://arxiv.org/abs/2609.32160 — arXiv 초기 증거 감사(초록) 2026-09-26

[188] https://arxiv.org/abs/2609.33209 — arXiv 확률 일관성 분석(초록) 2026-09-27

[189] https://simonwillison.net/2026/Sep/21/jev — Simon Willison's Weblog 2026-09-21

[190] https://raw.githubusercontent.com/browser-use/jev-ultrafast/main/docs/performance.md — Browser Use 성능 문서

[191] https://typesafe.ai/blog/introducing-system-one-models-and-jev — TypeSafe AI 발표 2026-09-15

[192] https://kode.kt.com/blog/article/8453 — KT kode 기술블로그(곽내원) 2025-10-29 — 상담Assist 지식추천 Agent LLM+RAG 구현

[193] https://m.corp.kt.com/archive/ipgrpt/attach/now/ESGFactbook_.pdf — KT ESG REPORT 2025 Factbook(문서 69쪽)

[194] https://corp.kt.com/attach/record/2026/KT%201Q26%20PT_SCRIPT_F.pdf — KT 2026년 1분기 실적발표 스크립트 2026-05-12

[195] https://www.nice.com/resources/agentic-ai-transforms-customer-experience-at-scale — NiCE/Cognigy Fabletics 고객 사례(발행일 미표기)

[196] https://www.genesys.com/company/newsroom/announcements/genesys-enhances-agentic-virtual-agent-amid-growing-enterprise-adoption — Genesys 2026-09-02

[197] https://www.bridgetec.co.kr/company/performance/domestic.jsp — 브리지텍 국내 구축 실적

[198] https://www.lgcns.com/kr/newsroom/press/detail.250825001 — LG CNS a:xink LG디스플레이 적용 2025-08-25

[199] https://help.nicecxone.com/content/copilotagents/copilotforagents.htm — NiCE Copilot for Agents 도움말