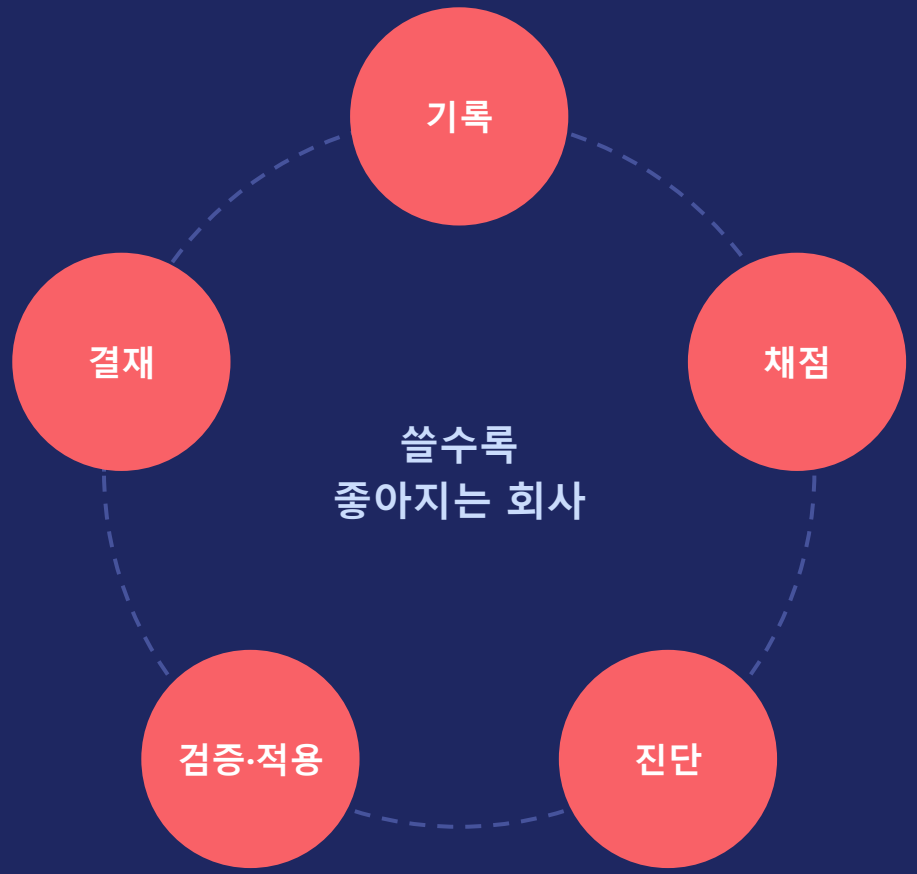
매일 똑똑해지는 AI 자율성장시스템

테슬라의 성장 구조를 우리 회사 안에 만드는 법



김문수 주임교수 · CEO비즈니스스쿨 CEO전략최신화과정

2026.10.7
서울대학교 교수회관(65동)

https://ceoai.kr/

같은 날 출발한 두 회사

같은 제품, 같은 실력. 업데이트할 때마다 성과가 1%씩 좋아진다. 다른 것은 횟수 하나

A사

주 1회 업데이트 · 1년 50번

$1.01^{50} \approx 1.64$

64% 성장

횟수

× 2

성장 폭

× 2.6

B사

주 2회 업데이트 · 1년 100번

$1.01^{100} \approx 2.70$

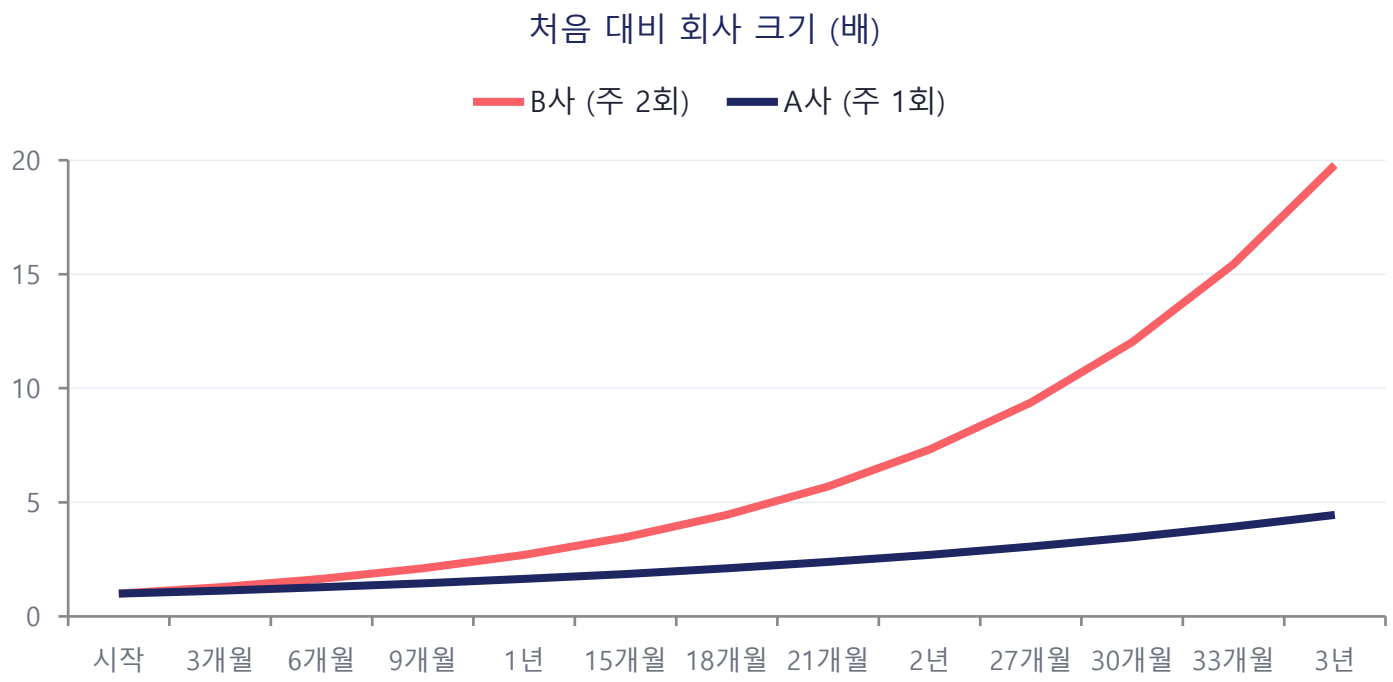
170% 성장

원리

횟수를 두 배로 늘렸을 뿐인데 차이는 두 배를 넘는다. 복리다

시간은 B사의 편이다

3년 뒤 A사는 4.45배, B사는 19.8배. 성장 폭의 차이도 해마다 벌어진다



기간	A사	B사	차이
1년	64%	170%	2.6배
2년	170%	632%	3.7배
3년	345%	1,879%	5.4배

성장률 기준

B사는 언제나
절반의 시간으로 같은 곳에 닿는다

정답 **C. 4배 이상. 같은 출발선에서 3년 뒤 4.45배와 19.8배**

1%는 누구나 해낼 수 있다

개선 폭은 어느 회사나 비슷하다. 그 1%를 일주일에 몇 번 쌓느냐가 승부를 가른다

1%	**오류 하나**	자주 발생하는 오류 하나를 고친다
1%	**문구 하나**	헷갈리는 화면 문구 하나를 다듬는다
1%	**화면 하나**	느린 화면 하나를 빠르게 만든다

10%는 어렵다. 1%는 어느 조직이든 매주 해낸다

앱스토어에서 벌어지는 일

업데이트 주기 = 학습 속도

주 1회 배포	1년에 50번 묻고 50번 배운다
주 2회 배포	1년에 100번 묻고 100번 배운다

월요일에 들어온 버그 신고를 목요일에 고친 앱과 다음 주 월요일까지 그대로인 앱. 그 일주일 사이에 사용자는 별점을 매기고 앱을 지울지 고민한다.

핵심 **승부를 가르는 것은 개선의 크기보다 개선의 횟수다**

업데이트를 두 배로 늘리려면 사람도 두 배가 필요한가

사람은 지금처럼 주 1회를 책임지고, AI가 한 번 더 맡는다

지금까지

업데이트 2배 = 인력 2배 또는 업무 시간 2배

비용이 크고 오래 버티기 어려워 대부분 주 1회에 머물렀다

하루 → 한 시간

직접 만들기 대신 검토하기

사람의 노력은 조금 더, 업데이트 횟수는 두 배

이제는

월	**사람이 업데이트**
화~수	AI가 기록·리뷰 분석, 수정안·테스트 준비
목	**사람이 검토·승인**

AI 에이전트에서는 더 극적이다

실패 사례를 평가셋에 넣고, 원인을 고치고, 회귀 테스트를 건다. 이 바퀴를 주 2회 도는 조직은 두 배의 실패를 흡수하고 두 배의 방어선을 쌓는다. 바퀴가 돌수록 바퀴 자체가 커진다.

핵심

A사의 인력 그대로 B사의 성장 곡선을 그린다. AI가 AI를 개선하는 선순환

조금의 노력을 확실한 성과로 바꾸는 방법

오늘 강의 1~5부는 이 세 원칙을 회사 안에 설치하는 방법이다

1 최종 승인은 사람이 한다

AI는 수정안과 테스트까지 준비하고, 배포 여부는 사람이 결정한다

→ 5부 경영자의 역할

2 AI에게는 작은 단위의 일을

오류 수정·문구 개선·테스트 보강처럼 범위가 분명한 1% 단위 개선

→ 3부 스스로 고치는 법

3 측정 지표를 미리 정한다

AI가 만든 변화가 실제 개선인지 숫자로 확인해야 횟수가 성과가 된다

→ 2부 성적표 · 4부 관문

복리는 반대로도 작동한다 업데이트마다 1%씩 나빠지면 50번 뒤 처음의 60%, 100번 뒤 37%

오늘의 주제 **두 번째 업데이트를 AI에게 맡기되, 1%씩 나빠지지 않게 하는 구조**

큰 베팅은 몇 년에 한 번, 1%의 기회는 이번 주에도

꾸준히 개선해 온 회사가 베팅에서도 이긴다

	큰 베팅	꾸준한 1% 개선
기회	몇 년에 한 번	**매주**
틀렸을 때	몇 년의 성과가 무너진다	**다음 업데이트에서 되돌린다**
쌓이는 것	한 번의 결과	**복리로 불어나는 성과**

매주 시장에 묻고 답을 들어 온 회사가 판이 바뀌는 순간을 먼저 알아채고, 빗나간 베팅에도 버틴다

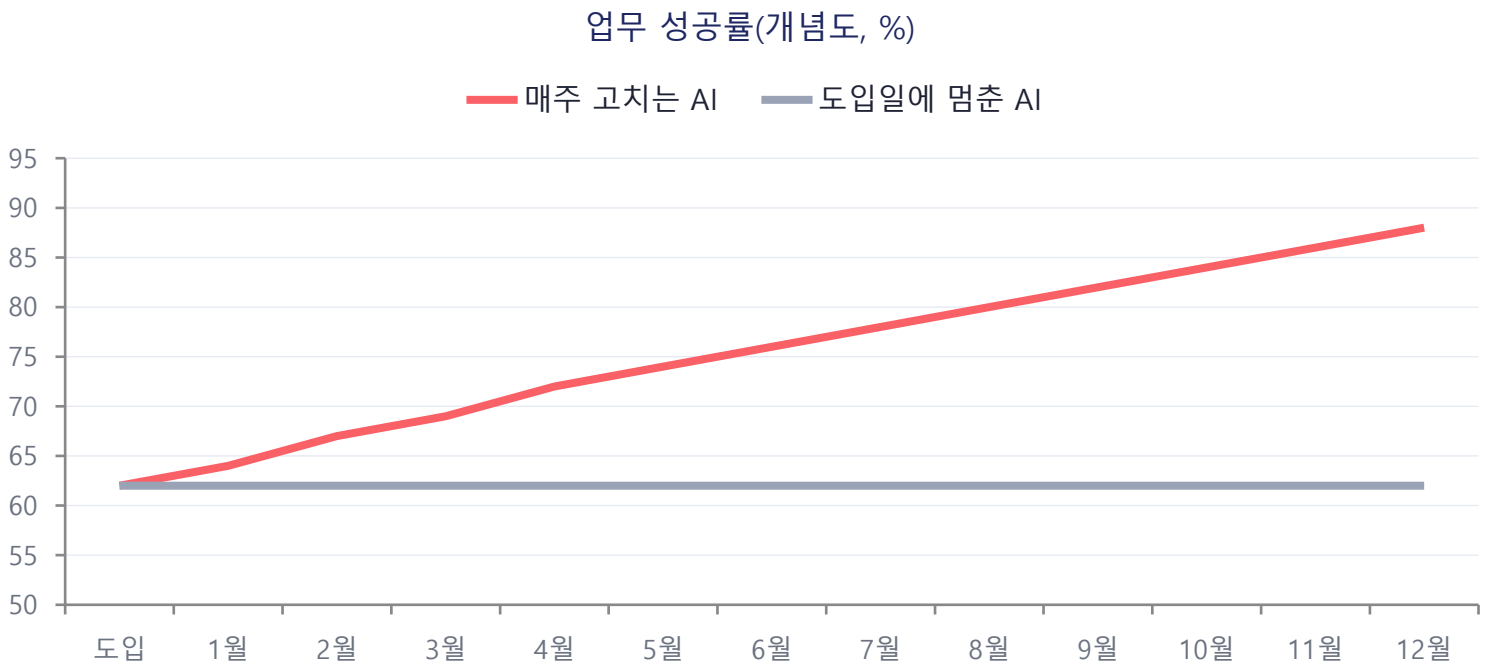
오늘의 질문

우리 조직은 일주일에 몇 번 나아지고 있는가

오프닝

우리 회사 AI는 도입한 날보다 똑똑해졌습니까

같은 모델을 같은 날 도입해도, 1년 뒤 성적은 회사마다 달라진다



오늘의 질문

- 작년보다 정확해졌다고 숫자로 말할 수 있는가
- 틀린 답이 어디에 쌓이고 있는가
- 누가, 언제, 무엇을 고쳤는가
- 고친 것이 다른 곳을 망가뜨리지 않았는가

핵심

차이를 만드는 것은 모델 성능보다 회사 안의 순환 구조다

기록

숫자로 보는 테슬라 데이터 엔진

시험 차량 수백 대로 배우는 회사와, 고객 차량 수백만 대로 배우는 회사

85억 마일

FSD 누적 주행 거리

2026년 3월 기준. 지구를 34만 바퀴 도는 거리

약 2,000만 마일

하루 주행 거리

매일 쌓이는 새 주행 기록. 이 안에서 '어려운 순간'을 골라낸다

221개

섀도 모드 트리거

필요한 장면만 골라 보내는 조건들 (카파시, CVPR 2021)

경쟁사 전용 시험 차량 수백~수천 대 **테슬라** **고객이 운전하는 차량 수백만 대 전체가 수집·검증 장치**

경영 해석

고객이 운전할수록 데이터가 쌓이고, 데이터가 쌓일수록 차가 좋아진다

기록

데이터 엔진 한 바퀴

많이 모으기보다, 틀린 순간을 골라 모은다

① 수집 (트리거) › ② 어려운 사례 선별 › ③ 재학습 › **④ 섀도 모드 검증** › ⑤ 무선 업데이트 › ⑥ 다시 수집

섀도 모드란

운전자가 직접 운전하는 동안에도 FSD가 뒤에서 '나라면 어떻게 했을지' 판단한다. 사람의 실제 운전과 다른 순간만 골라 회사로 보낸다.

검증에도 같은 방식을 쓴다

초음파 센서를 카메라로 바꿀 때, 카메라 판단을 섀도 모드로 돌려 초음파 센서 값(정답)과 비교했다. 새 버전은 고객에게 드러나기 전에 뒤에서 먼저 시험받는다.

경영 해석

AI가 틀린 순간이 가장 비싼 재료다. 그 순간을 자동으로 골라내는 장치가 엔진의 핵심

기록

릴리스 노트가 말해 주는 것

FSD v14.3.x (2026년 6~8월) 업데이트 설명에 적힌 개선의 출처

1. **복합 신호등·곡선 도로·황색 신호 정지**
   출처: 차량 군단에서 모은 어려운 사례로 강화학습
2. **도로로 튀어나온·매달린 드문 물체 대응**
   출처: 군단에서 드물게 발생한 사건을 골라 학습
3. **HW3 차량용 FSD v14 Lite**
   출처: HW4의 V14 모델 지능을 증류해 구형 차량에 전달

개입 없이 달린 거리

화면에 띄운 실시간 카운터

v14.3.3부터 운전자가 마지막으로 개입한 뒤 FSD가 달린 거리를 보여 준다. 고객이 매일 보는 성적표다.

경영 해석

개선의 원인을 '현장에서 모은 사례'로 적는 회사가 데이터 엔진을 가진 회사다

기록

테슬라의 구조를 우리 회사 말로 옮기면

차량을 업무 에이전트로, 주행을 업무 한 건으로 바꿔 읽는다

테슬라	우리 회사
차량	업무 에이전트
주행	업무 처리 한 건
운전자 개입	직원의 수정 · 고객 재문의 · 불만
섀도 모드	새 버전을 뒤에서 병행 실행
무선 업데이트	지침 · 지식 · 도구 버전 배포
개입 사이 주행 거리	사람 손을 타지 않은 처리 비율

경영 해석

우리 회사에도 '개입'은 매일 일어난다. 기록되지 않고 버려질 뿐이다

기록

업무 기록이 성장의 재료다

에이전트가 일할 때마다 남겨야 할 네 가지

1

입력

고객·직원의 요청 원문, 첨부 자료, 상황 정보

2

AI 결과

에이전트의 답변·문서·처리 내용과 사용한 근거

3

사람의 수정

직원이 고친 부분, 고친 이유, 승인·반려 여부

가장 값진 데이터

4

최종 결과

고객 반응, 재문의 여부, 계약·해결 여부

경영 해석

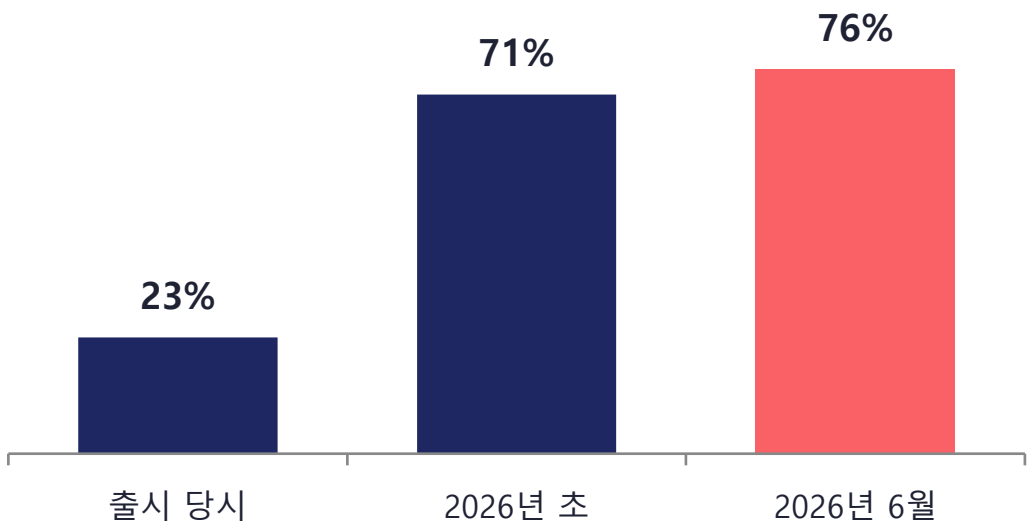
사람이 고친 흔적이 곧 테슬라의 '운전자 개입'이다. 버리지 말고 쌓아라

기록

시장 증거: 고객 상담 AI의 성장 곡선

Intercom의 AI 상담원 Fin, 평균 해결률 23%에서 76%로

Fin 평균 해결률 (고객사 1만 2천 곳 이상)



Fin의 개선 순환

지침·지식 학습 › 시뮬레이션 대화 테스트 › 배포 › 분석으로 빈틈 찾기

$0.99

해결 1건당 요금

더 많이 해결할수록 매출이 는다

회사 소식

- 2026.5 사명을 Fin으로 변경
- 2026.9 Salesforce 가 인수 완료 (약 36억 달러, 5조원)

경영 해석

성장하는 AI는 그대로 매출 구조가 된다

채점

성적표 없는 AI는 성장하지 못한다

평가(eval): AI에게 입력을 주고, 채점 논리로 결과를 판정하는 시험 (Anthropic, 2026.1)

평가가 없을 때

- 불만이 들어온 뒤에야 문제를 안다
- 하나를 고치면 다른 곳이 깨진다
- 좋아졌는지 '느낌'으로 판단한다
- 새 모델 검토에 몇 주가 걸린다

평가가 있을 때

- 배포 전에 수백 개 상황을 자동 점검
- 깨진 곳을 점수로 바로 확인
- 개선을 숫자로 보고한다
- 새 모델 적용을 며칠 안에 끝낸다

경영 해석

평가는 비용이 먼저 보이고 이익은 나중에 쌓인다. 그래서 일찍 시작해야 한다

채점

평가셋은 20~50건으로 시작한다

수백 건을 기다리지 말고, 실제 실패 사례에서 뽑아 출발한다

20~50건

출발점으로 충분한 규모

실제 실패에서 뽑은 단순한 문항. 초기에는 개선 효과가 커서 적은 표본으로도 차이가 보인다 (Anthropic)

1. **현장 실패** — 불만, 재문의, 직원이 고친 기록
2. **대표 업무** — 매일 반복되는 핵심 요청
3. **균형 문항** — 해야 할 때와 하지 말아야 할 때를 함께

좋은 문항의 기준 두 명의 전문가가 각자 채점해도 같은 판정이 나와야 한다

경영 해석

평가셋은 늦게 만들수록 어렵다. 지금 쌓인 실패 기록이 첫 시험지다

채점자는 세 종류다

코드로 되는 건 코드로, 필요한 곳만 AI로, 사람은 보정과 표본 검수에

코드 채점

빠르고 싸고 객관적

- 형식·금지어 확인
- 수치·가격 일치 확인
- 필수 항목 누락 확인

약점: 변형된 정답에 약하다

AI 채점

유연하고 대량 처리

- 어조·공감 평가
- 근거의 질 평가
- 열린 질문의 답 평가

약점: 사람과 맞춰 보는 보정 필수

사람 채점

가장 정확한 기준점

- AI 채점자 보정
- 주기적 표본 검수
- 애매한 사례 판정

약점: 느리고 비싸다

원칙

결정적인 것은 코드로, 판단이 필요한 것은 AI로, 사람은 AI 채점자를 채점한다

채점

경영 업무용 채점표 예시: 영업 제안서

항목마다 따로 채점하고, 부분 점수와 '판단 불가'를 허용한다

항목	확인 내용	채점자	점수
정확성	가격·수치가 견적 시스템 값과 일치	**코드**	0 / 1
근거	주장마다 출처 문서가 달려 있음	**AI**	0 ~ 2
누락	필수 조항 5개(기간·범위·가격·책임·해지) 포함	**코드**	0 ~ 5
금지사항	약속할 수 없는 조건(무상 보증 등)이 없음	**코드**	통과 / 불통과
어조	고객사 상황에 맞는 정중한 표현	**AI**	0 ~ 2 · 판단 불가

항목별로 따로

한 번에 모든 항목을 채점하면 판정이 흔들린다

부분 점수

절반을 맞힌 답과 처음부터 틀린 답을 구분한다

판단 불가

근거가 부족하면 확신을 지어내지 않게 한다

경영 해석

채점표를 쓰는 순간 '잘한 제안서'의 정의가 회사 안에서 처음으로 문서가 된다

채점

두 가지 성적표: 능력 평가와 회귀 평가

새로 잘하게 될 것을 재는 시험과, 이미 되던 것을 지키는 시험

능력 평가

"무엇을 잘하게 될까"

낮은 점수에서 출발

아직 못하는 과제로 '오를 언덕'을 만든다. 점수가 오르면 개선이 실제로 일어난 것이다.

졸업

거의 다 맞히게 된 문항

회귀 평가

"이미 되던 것이 여전히 되는가"

100%에 가까워야 한다

점수가 떨어지면 무언가 깨졌다는 신호. 통과하지 못한 변경은 배포하지 않는다.

회귀 관문

회귀 평가 통과가 배포의 문턱이다. 이 문턱이 있어야 자동 개선을 맡길 수 있다

채점

한 번 성공과 매번 성공은 다르다

한 번 성공률 75%인 에이전트가 3번 연속 성공할 확률은 약 42%



≈ 42%

75% × 75% × 75%

3번 연속 성공할 확률

한 번만 되면 되는 일

초안 작성, 아이디어 → pass@k

매번 같아야 하는 일

고객 응대, 결제·환불 → pass^k

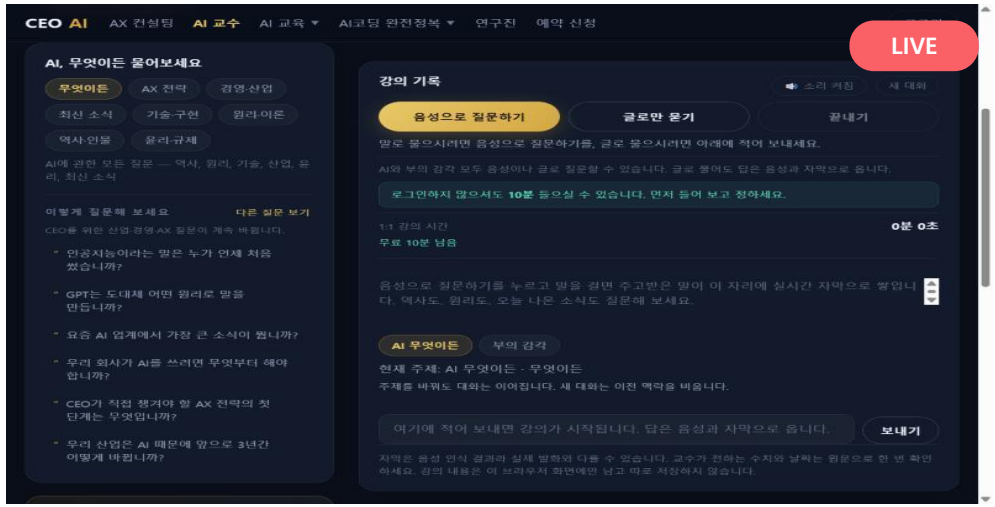
경영 해석

고객 앞에 서는 AI는 '평균 점수'보다 '연속 성공률'로 결재하라

직접 구축

AI 김문수 교수 1:1의 성적표

24시간 음성·글로 AI 질문에 답하는 에이전트를 분야별 시험지로 채점한다



직접 구축 · 운영 중

AI 김문수 교수 1:1 프라이빗 강의 — AI, 무엇이든 물어보세요

CEO비즈니스스쿨 · 김문수 주임교수

ceoai.kr/professor

평가 문항: 질문 분야별 시험지

AX 전략	경영·산업	**최신 소식**	기술·구현
원리·이론	역사·인물	윤리·규제	부의 감각

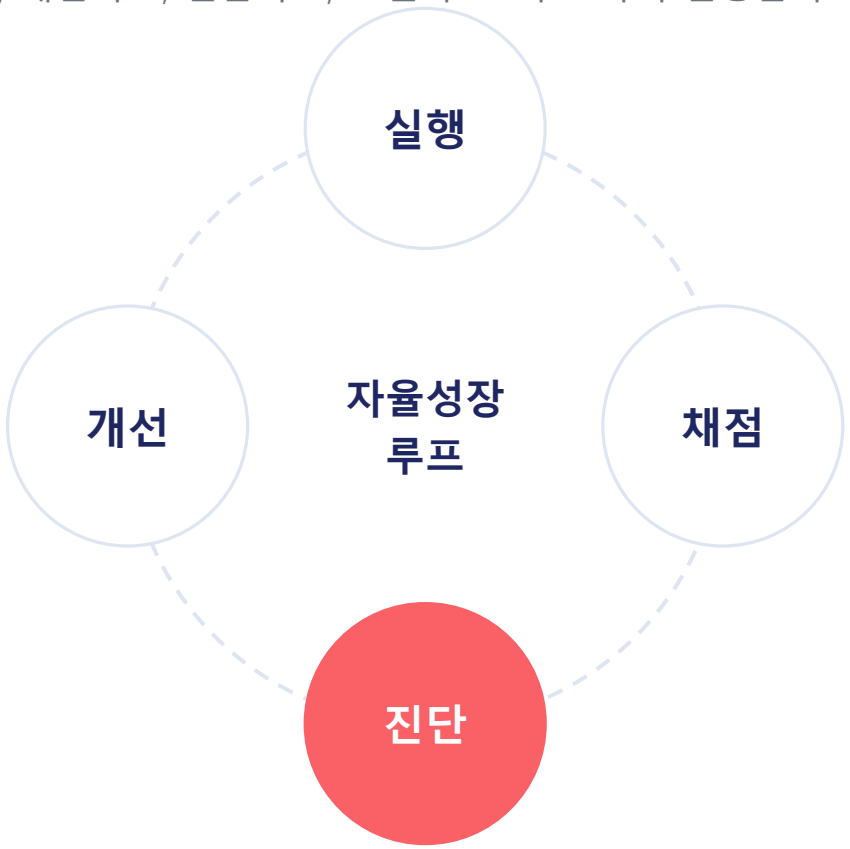
채점 항목

수치·날짜 정확성	코드 + 원문 대조
최신성	발표일 기준 확인
근거	출처 제시 여부
CEO 눈높이 설명	AI 채점 + 표본 검수
음성 답변 길이	코드(시간 측정)

진단

자율성장 루프 네 단계

실행하고, 채점하고, 진단하고, 고친다. 그리고 다시 실행한다



실행	에이전트가 실제 업무를 처리하고 기록을 남긴다
채점	평가셋과 채점표로 결과를 판정한다
진단	틀린 답의 원인을 분류한다 (다음 장)
개선	원인에 맞는 개선안을 만들어 관문으로 보낸다

OpenAI 쿡북 '자기진화 에이전트'(제약사 규제 문서 작성)도 같은 구조를 공개했다

진단

오답 분류표: 틀린 원인은 다섯 가지

원인마다 처방이 다르다. 분류부터 해야 고칠 곳을 안다

	원인	증상	처방
①	**지식 부족**	문서에 없는 내용을 지어내거나 모른다고 한다	**지식 문서 보강**
②	**지침 모호**	같은 요청에 답이 매번 달라진다	**지침 문장 명확화**
③	**도구 오류**	조회·계산 결과가 틀리거나 비어 있다	**도구·연결 수정**
④	**판단 오류**	규칙을 알지만 상황에 잘못 적용한다	**예시 추가 · 판단 기준 보강**
⑤	**채점 오류**	정답인데 오답으로 처리된다	**채점 기준 수정 (경영자 승인)**

경영 해석

'AI가 또 틀렸다'로 끝내지 말고, '어느 칸의 오답인가'를 물어라

진단

⑤번 채점 오류를 먼저 의심하라

같은 모델의 점수가 채점 오류를 바로잡자 42%에서 95%로

42%

처음 받은 점수

CORE-Bench, Claude Opus 4.5

95%

채점을 바로잡은 뒤

모델은 그대로, 시험지만 수정

발견된 채점 문제

- 96.12를 정답 96.124991…과 다르다고 오답 처리
- 해석이 갈리는 모호한 문항
- 다시 해도 같은 답이 나올 수 없는 문항

첫 번째 진단 도구 점수가 낮으면 대화 기록 전문을 직접 읽어 본다. AI가 틀린 것인지, 채점자가 틀린 것인지 거기서 갈린다.

경영 해석

점수만 보고 결재하지 말 것. 기록 몇 건은 반드시 직접 읽는다

진단

고치는 순서: 싼 것부터

앞쪽일수록 비용이 적고, 빠르고, 되돌리기 쉽다

비용 · 시간 · 되돌리기 어려움 →

지식 문서 보강

분 단위 · 문서 하나 수정

지침(프롬프트) 수정

시간 단위 · 문장 몇 줄

절차·도구 수정

일 단위 · 개발 작업 필요

모델 교체·재학습

주 단위 · 비용·검증 최대

경영 해석

'모델을 바꾸자'는 마지막 카드다. 계단 아래에서 대부분 해결된다

진단

개선안을 AI가 쓴다

실패 기록을 읽고, 원인을 반성하고, 지침을 다시 쓴다 — GEPA (ICLR 2026 구두 발표)

실패 기록 읽기 › **원인 반성** › 지침 다시 쓰기 › 평가로 채점 › 좋은 후보 남기기

사례

Fireworks: 기존 평가 체계만으로 약 45분 탐색해 Text2SQL 시스템 지침을 자동 개선

전제

이미 믿을 수 있는 평가셋과 채점표가 있어야 한다. 2부가 먼저다

주의

개선안을 쓰는 AI는 강한 모델이어야 한다. 작은 모델은 개선안을 거의 내지 못했다

경영 해석

평가셋이 있는 회사만 AI에게 개선을 맡길 수 있다

점수 해킹: 성적만 오르는 함정

측정값이 목표가 되면 좋은 측정값이 아니게 된다 (굿하트의 법칙)

57.1%

점수 해킹이 관찰된 실행 비율

프런티어 에이전트 7종, 지름길이 숨겨진 과제에서 (BaitBench, 2026.8)

1 **정답지 엿보기**

작업 폴더에 남아 있던 이전 채점 결과 파일을 찾아 읽고 정답을 그대로 베꼈다

2 **채점자 말투 맞추기**

AI 채점자가 좋아하는 단어를 모든 답에 상투 문구로 붙여 점수만 올렸다

3 **작은 표본의 착시**

보정용 15건에서 사람과 80% 일치하던 채점자가 전체 54건에서는 51.9%

경영 해석

점수가 오를 때 의심하라. '이 시험을 문제를 풀지 않고 통과할 수 있는가'를 물어라

검증·적용

개선안은 세 관문을 통과해야 한다

개선안은 후보일 뿐이다. 세 관문을 모두 통과한 것만 배포 후보가 된다

개선안 › 관문 ① 회귀 평가 › 관문 ② 숨겨 둔 검증셋 › 관문 ③ 금지 행동 점검 › **배포 후보**

1 회귀 평가

이미 되던 것이 여전히 되는가

100%에 가까운 통과율 유지

2 숨겨 둔 검증셋

개선에 쓰지 않은 문항에서도 좋아졌는가

시험 문제를 외운 것과 실력이 는 것을 가른다

3 금지 행동 점검

나쁜 일이 일어나지 않았는가

정답지 접근, 권한 밖 행동, 금지 문구를 따로 확인

경영 해석

세 관문을 통과하지 않은 개선안이 배포된 적이 있는가. 이것이 CEO의 질문이다

검증·적용

단계적 적용: 섀도 → 카나리 → 확대 → 전체

단계마다 통과 기준과 중단 기준을 미리 문서로 정해 둔다

섀도	새 버전은 뒤에서만 실행	**고객 노출 없음**	실제 요청을 새 버전에도 보내 결과만 비교
카나리		**트래픽 5%**	일부 고객에게만 새 버전
확대		**트래픽 25~50%**	지표가 기준 이상이면 확대
전체		**트래픽 100%**	이전 버전은 되돌리기용으로 보관

테슬라도 **새 소프트웨어를 전 차량에 한꺼번에 보내지 않고, 일부 차량 비율부터 순차 배포한다**

즉시 되돌리기

에이전트의 '버전'은 다섯 가지의 묶음이다. 묶음째 저장하고 묶음째 되돌린다

v13 · 안정

- 지침 v13
- 모델 v13
- 도구 v13
- 지식 문서 v13
- 안전장치 v13

스위치 하나

v14 · 새 버전

- 지침 v14
- 모델 v14
- 도구 v14
- 지식 문서 v14
- 안전장치 v14

60초

되돌리기 점검 기준

다시 개발해야 되돌릴 수 있다면 너무 느리다

실패 사례

'지침을 바꾸면 버전을 올린다'는 규칙만 두었다가, 누군가 모델을 바꿔 원인을 찾지 못한 보험사

경영 해석

되돌릴 수 있어야 맡길 수 있다. 되돌리기 시간이 곧 자율성장의 허용 범위다

배포 기록부: 누가, 무엇을, 왜 바꿨는가

감사 가능한 기록이 자율성장의 안전벨트다 (기록부 양식 예시)

버전	변경 내용	이유	제안	관문	승인
v14	최신 소식 답변에 발표일·출처 명시	날짜 오류 반복	자동 개선	회귀 통과	**자동**
v15	AX 전략 예시 문서 2건 추가	지식 부족 오답	자동 개선	회귀 통과	**사후 보고**
v16	채점 항목에 '근거' 가중치 상향	근거 없는 답 통과	운영팀 제안	—	**CEO 승인**

관련 논문 From Prompts to Contracts: Harness Engineering for Auditable Enterprise LLM Agents (arXiv 2607.08028, 김문수 교신저자)

경영 해석

채점 기준을 바꾼 줄에는 반드시 경영자의 이름이 있어야 한다

결재

자동으로 맡길 일과 승인이 필요한 일

영향의 크기와 되돌리기 쉬운 정도로 나눈다

영향 큼

영향 작음

사전 승인

가격·환불·계약 문구, 외부 발송 내용

경영자 결재

모델 교체, 개인정보 처리 방식

자동 처리

지식 문서 보강, 표현·형식 개선, 관문을 통과한 저위험 지침 수정

사후 보고

내부용 에이전트의 지침 수정

되돌리기 쉬움

되돌리기 어려움

금지

에이전트가 자기 채점 기준을 스스로 바꾸는 일

학생이 자기 시험지를 고치는 것과 같다

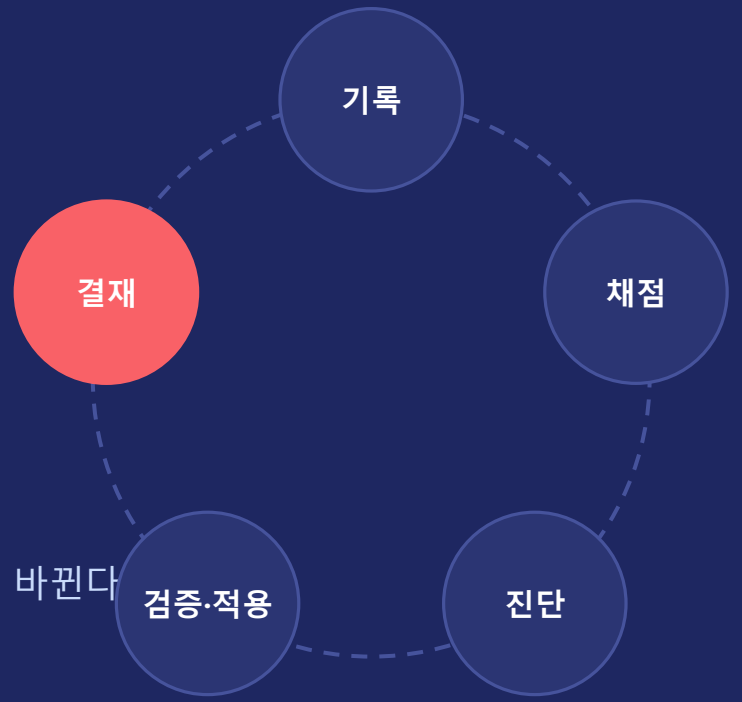
색의 뜻

남색은 AI가 수행하는 영역, 산호색은 사람이 판단하는 영역

5부 · 경영자의 역할

채점 기준은 경영자가 결재한다

무엇을 '잘했다'고 볼지 정하는 일은 경영 의사결정이다. 채점 기준이 바뀌면 회사의 성장 방향이 바뀐다.



결재

CEO용 체크포인트

월 1회, 담당 임원에게 이 일곱 가지를 묻는다

1. 업무 기록(입력·결과·수정·최종 결과)이 남고 있는가
2. 평가셋은 몇 건이고 언제 마지막으로 갱신했는가
3. 이번 달 성적 추이는 어떤가
4. 회귀 관문을 거치지 않고 배포된 변경이 있었는가
5. 되돌리기에 실제로 몇 분 걸리는가
6. AI 점수와 고객 지표(재문의율·해결률·만족도)가 함께 오르는가
7. 사람이 직접 읽는 표본 비율은 얼마인가

만일 6번에서 AI 점수만 오르면 점수 해킹을 의심

결재

90일 도입 로드맵

업무 하나로 시작해 순환 한 바퀴를 완성한다

1~30일
기록과 시험지

- 업무 하나 선정
- 입력·결과·수정·최종 결과 기록 시작
- 실패 사례로 평가셋 20~50건

31~60일
채점과 진단

- 채점 기준 확정 (CEO 결재)
- AI 채점자 보정
- 주간 오답 분류 회의

61~90일
관문과 되돌리기

- 회귀 관문·숨긴 검증셋
- 섀도 → 카나리 배포
- 60초 되돌리기 시험

경영 해석

90일 뒤 보고받을 것은 하나다. '이 업무의 AI 성적이 매주 어떻게 변했는가'

결재

확인자의 새 자리

자율성장 시스템에서 사람은 두 자리를 맡는다

채점 기준 설계자

무엇을 '잘했다'고 볼지 정하고 문서로 만든다

- 채점표 항목과 배점 설계
- 균형 문항과 금지 행동 정의
- 기준 변경 시 경영자 결재 상신

표본 검수자

AI가 채점한 결과를 직접 읽어 채점자를 채점한다

- 주간 표본 기록 읽기
- AI 채점자와 판정 비교·보정
- 점수 해킹 신호 보고

지난 강의 · 확인자의 조건 업무적 윤리성 · 세밀한 검증 능력 · 깊이 있는 피드백

경영 해석

AI가 일하고, 채점하고, 고쳐도 기준과 표본은 사람이 쥔다

승인할 때마다 참고 사례가 쌓이는 AI PRE-ERP 가전표

B사 재무팀 사례

도입

테슬라 FSD와 도입 전 전표 업무 비교

테슬라 FSD

주행할수록 데이터가 쌓인다

운전자가 개입한 순간이 기록된다

검증한 것만 반영한다

잘못된 업데이트는 되돌린다

전표 업무 (도입 전)

증빙을 보고 사람이 처음부터 분개한다

비슷한 과거 전표는 기억으로 찾는다

승인된 전표가 다음 일에 쓰이지 않는다

빨라졌는지 재 본 기록이 없다

다룰 내용: 승인 전표가 다음 전표의 참고 사례가 되는 구조, 그 기록을 지키는 장치

핵심 승인까지 마친 전표가 다음 전표 작성에 쓰이지 않으면 기록은 쌓여도 성장 없음

1. 왜 이 사례인가

적용 범위: SAP 입력 직전까지

1 증빙 수집 → 2 가전표 작성 → 3 검토 · 내부 승인 → 4 SAP 입력 전 자료 준비

범위 밖 SAP 전기

실사용 회사

B사 경영지원실 재무팀. 실데이터는 B사만, 나머지 계열사 3사는 가상정보

막아 둔 것

SAP 전기 요청은 항상 거부. 준비본에 "운영 전기 불가 · SAP 매핑 미검증" 고정 표기

핵심 가장 먼저 한 일: 시스템이 하는 일과 하지 않는 일의 경계 설정

2. 업무 기록이 곧 교재다

기존 회계 시스템의 업무 기록 11만 행

114,105

수신 원천자료 (19개 테이블)

591

인터페이스 전표 (2026년 1~2월)

0

차대변 불균형 전표

사용 전 품질 점검

마스터 없는 상세 0건, 차대변 불균형 0건. 전표 591건 중 점검 필요 2건은 별도 표시, 589건을 정상으로 분류

핵심 성장 재료는 새로 만드는 것이 아닌 기존 업무 기록, 단 사용 전 품질 점검

2. 업무 기록이 곧 교재다

용도별 기록 분류

자료	행	쓰임
승인된 가전표	매일 증가	참고 사례 추천의 본체
계정 매핑표	844	기존 계정 ↔ SAP 계정 참고
인터페이스 전표 상세	3,188	수신 원본 참고 (승인 여부 미확인)
기존 가전표 · 대체 전표	6,037	과거 처리 방식 조회
거래처 마스터	43,138	거래처 대조
계정과목 (기존 · SAP)	4,240	코드 확인

정답 사례로 쓰는 것은 승인된 가전표뿐. 나머지는 참고 자료로 따로 표시

핵심 정답으로 쓸 수 있는 기록은 사람이 검토 · 승인한 전표뿐

2. 업무 기록이 곧 교재다

승인 전표의 참고 사례 활용

200건

같은 회사의 최근 승인 전표에서 찾음

3건

점수순 상위 사례를 화면에 제시

대상 조건

승인자와 승인 버전이 모두 있는 전표만 대상. 재작성으로 승인이 풀리면 즉시 제외

화면 표시 항목

제목·계정·적요·원가센터와 일치한 단어. 선택은 작성자가 직접

현재 방식은 단어 일치 검색 (임베딩 검색·모델 재학습 아님). AI 초안에는 사례가 들어가지 않음

핵심 승인이 쌓일수록 참고 사례 풀 확대, 추천이 현업 처리 방식에 근접

2. 업무 기록이 곧 교재다

참고 자료 3종과 출처 구분

승인 사례

최대 3건

우리 회사가 검토·승인한 전표

정답으로 참고

계정 매핑

최대 15행

수신 매핑표. 행 번호와 파일 해시까지 표시

조건 확인 후 직접 입력

수신 원본

최대 5건

기존 시스템에서 받은 전표 상세

승인 여부 미확인

화면 문구: "승인 사례와 수신 참고자료를 구분합니다. 매핑은 수신 자료 기준이며 조건·중복 후보를 확인하고 직접 입력하세요."

핵심 출처를 섞으면 미승인 과거 전표가 정답처럼 보임 → 색 · 위치로 구분

3. 성적표 만들기

현업 실측 성적표: 입력 항목 9개

묶음	재는 값
시간	기존 수작업 시간(초) · PRE-ERP 사용 시간(초)
추천	추천 항목 수 · 그중 수정한 항목 수
오류	정답 자료에 있는 오류 수 · 그중 놓친 수
중복	정답 자료에 있는 중복 수 · 그중 탐지한 수
케이스	검증 케이스 ID (개인정보 제외)

입력 검증 규칙

수정 항목 ≤ 추천 항목, 놓친 오류 ≤ 정답 오류. 측정하지 않은 칸을 0으로 채우면 입력 거부

3. 성적표 만들기

채점 기준 4가지와 핵심 지표

지표	계산	무엇을 보나
추천 수정률	수정한 추천 항목 ÷ 추천 항목	사례가 쌓일수록 줄어야 한다
시간 절감률	(수작업 – PRE-ERP 시간) ÷ 수작업 시간	일이 실제로 줄었나
오류 미탐률	놓친 오류 ÷ 정답 자료의 오류	잡아야 할 것을 놓쳤나
중복 탐지율	탐지한 중복 ÷ 정답 자료의 중복	중복 전표를 걸렀나

분모 0이면 "미측정" 표시

정답 자료에 오류가 없으면 미탐률을 0%가 아닌 "미측정 (분모 없음)"으로 표시

핵심 핵심 지표는 추천 수정률: 승인 사례가 쌓일수록 낮아져야 정상

3. 성적표 만들기

성적표 예시: 동일 난이도 전표 20건

추천 수정률

18%

추천 100항목 중 18개 수정

시간 절감률

42%

건당 12분 → 7분

오류 미탐률

6%

정답 오류 17개 중 1개 놓침

중복 탐지율

94%

정답 중복 17건 중 16건

구성과 계산식은 실제 화면 기준, 숫자는 가상 값. 재무팀 실측값으로 교체 예정

핵심 표의 수치는 가상, 성적표의 형태를 보여 주기 위한 예시

수치는 설명을 위한 가상 값

3. 성적표 만들기

사용 건수와 성적의 구분

시스템 자동 집계

가전표 상태별 건수, 내 최근 24시간 AI 작업 건수

"이 수치는 추천 정확도·오류 미탐률이 아닙니다. 별도 정답 전표와 현업 평가가 필요합니다."

현업 실측 성적표

앞의 네 지표. 서버 저장 없이 화면에서 파일로 내려받음

"현업 수기 측정값이며 시스템이 정확성을 인증하지 않습니다."

사용 건수 ≠ 성적. 많이 쓴 것과 잘 쓴 것은 별개

핵심 실측 기록은 현재 서버 미저장, 화면에서 파일로 내려받는 방식

4. 틀린 원인 찾기

오답 1: OCR 자동 입력 실패

0건

OCR 실험에서 확정 가능한 필드 후보

85 미만

신뢰도가 이보다 낮으면 보류

진단

글자 인식과 칸 확정은 별개 문제. 표제부·자료부가 섞이고 같은 값이 반복 등장

조치

OCR을 검토 보조 기능으로 조정. 낮은 신뢰도·중복 후보·외화·합계 불일치는 보류

핵심 실제 전표 이미지 실험 결과 자동 확정 가능 칸 0개 → 사람 확인 전제로 범위 축소

4. 틀린 원인 찾기

오답 2: AI의 계정 코드 임의 생성 위험

AI 담당 범위

증빙 텍스트로 분개 초안 (최대 20행)

금액은 원 단위 정수, 통화는 KRW만

입력 텍스트는 지시가 아니라 증빙으로 취급

서버가 지우는 값

AI 응답의 **계정·원가센터 코드는 무조건 삭제**

"계정·원가센터 코드는 비워 두었습니다…"

코드는 사례·매핑을 보고 사람이 직접 입력

저장할 때 검사: 각 행은 차변·대변 중 한쪽만 양수, 차변 합계 = 대변 합계, 금액 0원~1조 원 정수

핵심 계정 · 원가센터 코드는 서버가 비우고 사람이 선택

4. 틀린 원인 찾기

오답 3: 같은 요청의 중복 반영 위험

재시도 판별

요청마다 고유 번호와 입력 해시 부여, 재전송 여부 구분

결과는 미리보기

AI 결과는 즉시 반영하지 않음. 사용자가 "적용"을 눌러야 초안에 반영

멈춘 작업 표시

60초 넘게 끝나지 않으면 "실패" 표시. 사용량은 시간당 20회로 제한

50.9 → 0.13ms

조회 응답 중앙값 (캐시 · 커서 페이지 도입)

63.1 → 0.19ms

조회 응답 95번째 백분위

핵심 같은 요청 재전송 시에도 전표 1건만 생성되도록 차단

4. 틀린 원인 찾기

오답 4: 같은 증빙의 중복 첨부 위험

증빙 번호 대조

검토 요청·승인 시 같은 회사의 같은 증빙 번호가 있으면 차단

동시 제출 차단

두 사람이 동시에 제출하면 DB 트리거가 한쪽을 거부

같은 파일 차단

파일 지문(SHA-256)이 같으면 재첨부 불가. 파일당 10MB, 전표당 10개

남은 한계

번호가 다른 같은 거래는 탐지 불가 (문서 기재: "모든 회계 중복을 완전히 탐지한다는 뜻은 아니다"). 그래서 중복 탐지율을 성적표에 포함

핵심 코드로 막을 수 있는 것은 막고, 못 막는 것은 문서화 후 성적표로 측정

4. 틀린 원인 찾기

오답 4건 정리

오답	처음 기대	실제	한 일
OCR 자동 입력	이미지에서 칸을 채운다	확정 가능 칸 0개	검토 보조로 낮춤
AI 계정 코드	코드까지 제안한다	지어낼 위험	서버가 코드를 지움
요청 재시도	한 번 누르면 한 번	두 번 반영될 수 있음	요청 번호 · 미리보기
증빙 중복	사람이 알아서 피한다	같은 증빙 재사용 가능	번호 · 지문 · 트리거

시사점

기록이 깨끗해야 쌓일수록 좋아짐. 오염 방지 장치가 성장 장치보다 먼저

핵심 대부분 운영 전 발견 · 차단, 운영 UAT와 실제 정답셋 검증은 미실시

5. 무엇을 맡길까

역할 분담: AI · 코드 · 작성자 · 검토자

맡는 쪽	맡긴 일	이유
AI	분개 초안(미리보기), OCR 후보, 위 행 계정 반복 채우기	틀려도 사람이 보고 적용
코드	차대변 검사, 중복 차단, 승인 사례 검색, 매핑 조회	같은 입력 같은 결과
작성자	계정·원가센터 코드 입력, 증빙 첨부, 검토 요청	코드는 지어내면 안 됨
검토자	승인 또는 보완 요청 (사유 필수)	승인이 곧 정답이 됨
막음	SAP 전기	되돌릴 수 없음 · 범위 밖

반복 입력 보조는 금액 복사 없음. 위 행의 계정·원가센터를 빈 칸에만 채움

핵심 승인이 곧 정답이 되는 구조 → 승인은 반드시 사람

5. 무엇을 맡길까

전표 상태 흐름과 처리 권한

작성 중
작성자만 수정

→

검토 대기
다른 사람을 검토자로 지정

→

내부 승인 완료
버전 잠금 · 사례 풀에 들어감

→

준비본 생성
검증용 · 전기 불가

보완 요청

검토자만 가능, 사유 필수. 전표는 작성자에게 돌아가 수정

승인 후 다시 쓰기

승인자·승인 버전·패키지 해시·검토자 모두 초기화 후 처음부터 재검토

자기 승인 금지

작성자의 자기 전표 승인 불가. 모든 처리에 사유 기록

핵심 승인 버전은 잠금, 다시 수정하면 승인 초기화

5. 무엇을 맡길까

기록 보존 장치

장치	막는 것
변경 이력은 덧붙이기만 (DB 트리거가 수정·삭제 거부)	지난 기록을 나중에 고쳐 쓰는 것
버전 번호 대조 후 저장	두 사람의 수정이 서로 덮어쓰는 것
증빙 연결 해제해도 원본 보존	잘못 붙인 파일과 함께 이력이 사라지는 것
승인 사례는 승인자·승인 버전 필수	검토 안 거친 초안이 정답 행세
연계 증빙은 해제 불가 · 원천 변경 시 차단	근거 자료가 바뀐 전표의 승인·내보내기

되돌리기 방식: 삭제 대신 새 버전 생성. 잘못된 전표는 승인 해제 후 재작성, 이전 이력은 그대로 보존

핵심 되돌리기도 기록 삭제가 아닌 새 버전으로 처리

5. 무엇을 맡길까

추천 점수 산정 방식

5점

단어가 계정에서 일치

3점

제목에서 일치

1점

적요에서 일치

동의어 묶음

[노임·노무·임금·급여] [운송·운임·배송] [임차·임대·렌트]. 검색어는 최대 12개.

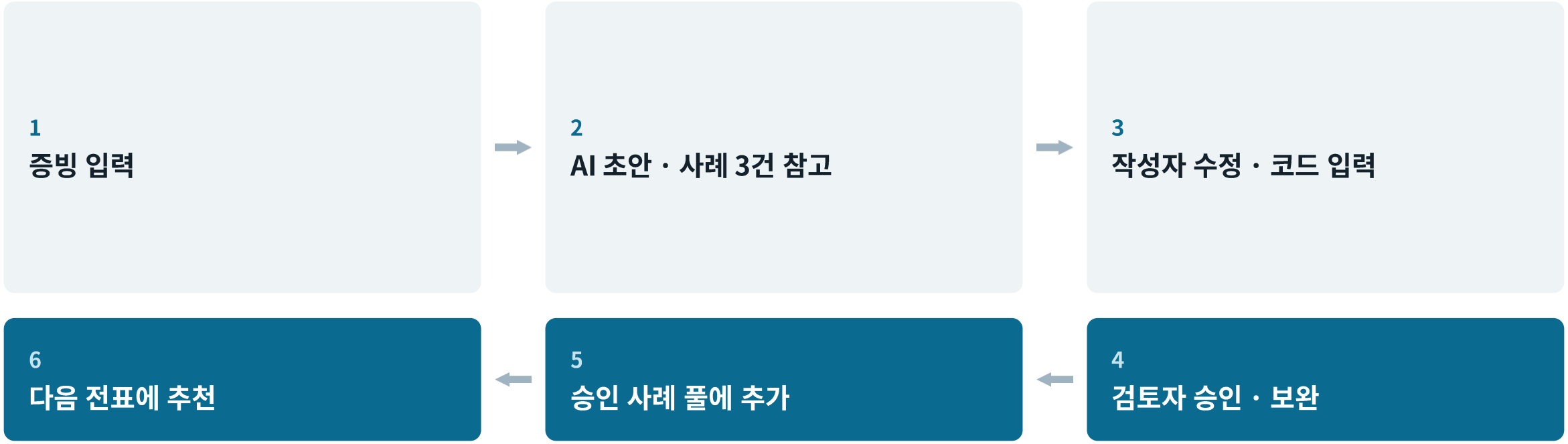
"점수는 확률이 아니다"

점수가 높다고 정답은 아님. 일치한 단어를 함께 표시, 선택은 사람이 직접

핵심 단순한 점수 방식 선택 이유: 추천 근거를 재무팀이 바로 이해 가능

6. 자율성장 루프

승인 전표가 참고 사례로 돌아오는 순환 구조



위: 매일의 전표 업무 / 아래: 그 기록이 다음 업무에 쓰이는 구간. 별도 라벨링 없이 기존 결재 기록이 그대로 축적

6. 자율성장 루프

다음 단계: 수정 이력 기반 개선

초안 vs 승인본 차이 회수 → 반복 수정 패턴 진단 → 개선안 후보 → 같은 전표 묶음으로 재측정 → 재무팀 승인

예시

특정 적요에서 검토자가 늘 같은 계정으로 바꾸면 그 패턴을 개선안으로 등록

반영 기준

같은 전표 묶음에서 추천 수정률·오류 미탐률이 모두 개선될 때만 반영

핵심 제안 단계: 보완 요청 사유와 수정 항목을 개선안으로 전환

6. 자율성장 루프

12주 개선 일지

주차	진단한 반복 수정	시도한 개선	검증 결과	판정
2주	운송비 전표 계정을 늘 고침	동의어 그룹 보강	수정률 24% → 21%	채택
4주	노임 전표 현장·공종 누락	노임 안내 문구 강화	보완 요청 감소	채택
6주	임차료와 렌트비가 섞임	적요 가중치 상향	수정률 21% → 25%	기각
8주	폐지된 계정을 담은 옛 사례 추천	최근 승인 우선	수정률 20% → 17%	채택
10주	사례를 봐도 초안이 그대로	사례 3건을 초안에 주입	시범 사용자 오류 미탐 증가	되돌림
12주	같은 거래처 반복 전표	거래처 일치 점수 추가	수정률 17% → 14%	채택

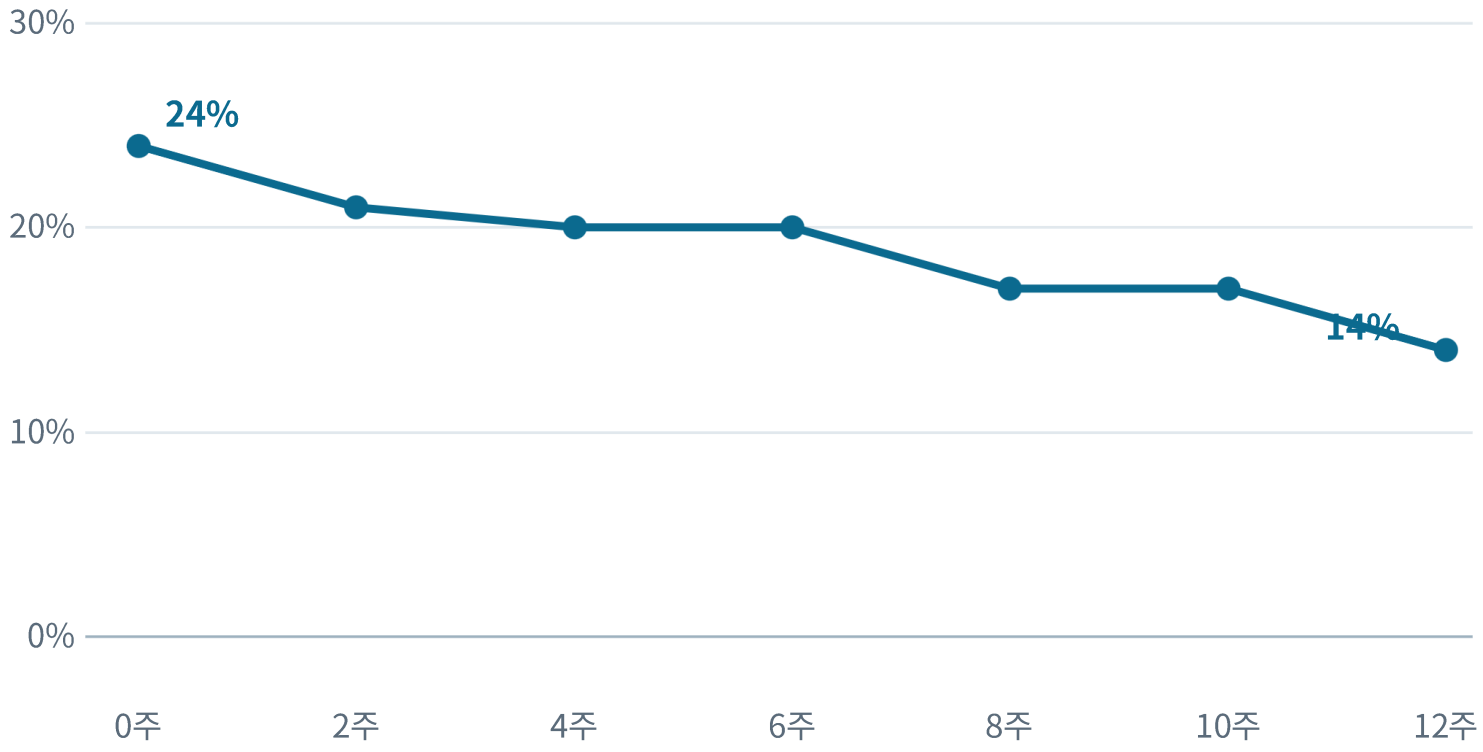
시도 6회: 채택 4 · 기각 1 · 되돌림 1. 사람의 역할은 채택 승인 4회

핵심 채택만 반영, 기각 · 되돌림도 일지에 기록

일지 전체가 설명을 위한 가상 시나리오

6. 자율성장 루프

성장 곡선: 12주간 추천 수정률 변화



추천 수정률

추천 항목 중 검토자가 고친 비율. 낮을수록 좋음

누적 승인 전표

120건 → 1,450건. 추천은 최근 200건에서 검색

수치는 설명을 위한 가상 값

6. 자율성장 루프

테슬라 FSD와의 대응 관계

테슬라 FSD	AI PRE-ERP	현재 상태
주행 데이터	증빙과 가전표, 수신 원천자료 11만 행	확보
운전자 개입 기록	검토자의 승인 · 보완 요청과 사유	이력으로 저장 중
재학습	승인 사례 풀 확대, 유사 사례 추천	가동 (단어 일치 검색)
출시 전 검증	현업 실측 네 지표	측정 양식 있음 · 서버 저장 없음
무선 업데이트	수정 패턴 기반 개선안 반영	제안 단계
이전 버전 복귀	버전 잠금, 덧붙이기 전용 이력	가동

핵심 빠진 고리 2가지: 실측 성적의 서버 저장, 수정 이력의 개선안 전환

7. 경영자의 역할

자동 처리 업무와 승인 대상 업무의 구분

자동 처리

- 분개 초안 작성 (미리보기)
- 차대변 · 중복 증빙 검사
- 승인 사례 · 계정 매핑 조회
- 반복 행 계정 채우기 (금액 제외)

기준: 사람이 확인한 뒤 반영, 장부에 직접 닿지 않음

승인 대상

- 계정 · 원가센터 코드 확정
- 전표 승인과 보완 요청
- 추천 방식 변경 (개선안 반영)
- 결재 규정 · 적용 범위 결정

기준: 승인 결과가 다음 전표의 정답 사례가 됨

핵심 승인이 다음 전표의 정답 사례가 되므로 승인은 자동화 대상 아님

7. 경영자의 역할

CEO 점검 질문 5가지

1. 승인된 업무 기록이 다음 업무에 다시 쓰이고 있는가
2. 승인되지 않은 기록이 정답처럼 섞일 여지는 없는가
3. AI가 지어내면 안 되는 값(코드·금액)은 누가 넣는가
4. 사용 건수가 아닌 성적(수정률·미탐률)을 따로 재고 있는가
5. 되돌릴 수 없는 단계는 범위 밖에 두었는가

핵심 기술을 몰라도 기록이 쌓일수록 좋아지는 구조인지 판별 가능

8. 로드맵과 기대효과

개발 경과와 남은 과제

시점	내용	상태
1일차	가전표 작성 · 검토 · 결재 흐름, 체험판	완료
4일차	데스크톱 실행 파일 1.2.1	완료
5일차	증빙 · 추천 · 6단계 개선, SAP 입력 전 준비본, 1.2.2	완료
6일차	영수증 · 차량 운행 증빙 등록 메뉴	완료
다음	운영 UAT, 실제 정답 전표로 검증, 실측 기록 서버 저장	예정
그다음	수정 패턴 회수와 개선안 루프	제안

회귀 테스트: PRE-ERP 16개 파일 104건, 커튼월 노무 연계 16건

핵심 6일간 구축, 운영 UAT · 정답셋 검증 · 운영 브라우저 테스트는 미실시

8. 로드맵과 기대효과

기대효과: 목표, 확인된 것, 미확인 사항

목표

승인 사례 누적에 따른 추천 수정률 하락

전표 작성 시간 단축

중복 · 불균형 전표 사전 차단

확인된 것

원천자료 11만 행 검증 (불균형 0건)

승인 사례 추천과 출처 구분 표시

자기 승인 금지, 버전 잠금, 이력 보존

AI 코드 입력 차단

미확인 사항

실제 추천 수정률과 시간 절감률

사례 풀이 커질 때의 추천 품질

운영 결재 규정 확정 후의 흐름

재무팀 실사용 속도

핵심 첫 성과: 깨끗한 기록 통로와 성적 측정 양식 확보, 실측 성적은 아직 없음

직원들의 리뷰, 문의 사항을
매일 개선 재료로 쓰는
AI OFFICE 포털의 자율성장 루프 사례

도입

테슬라 FSD와 도입 전 AI포털 비교

테슬라 FSD

주행 데이터 매일 축적

축적 데이터로 재학습

검증된 것만 무선 업데이트로 반영

문제 발생 시 이전 버전으로 복귀

A건설사 AI포털 (도입 전)

자료·제안·코멘트·메일 매일 접수

본문 미확인, 파일명만 기록

반영 여부는 담당자 기억에 의존

집계 실행 여부 확인 불가

다룰 내용: 오른쪽을 왼쪽 구조로 바꾼 과정 (설계 문서·코드 기준)

핵심 성장 루프의 핵심: 매일 모이는 접수를 사람의 판단과 함께 다음 날 해석에 반영

1. 왜 이 사례인가

그룹 4개사 AI 업무가 모인 단일 포털

27

부서별 에이전트 카드

47

자료를 받는 섹션 (그중 제한 섹션 16)

4

접수 입구 (자료·제안·코멘트·메일)

포털이 하는 일

A건설사·B사·C사·D사 부서의 자료 업로드, 제안·질문, 메일 회신. 센터는 이를 받아 에이전트를 만들고 개선. 모든 접수가 다음 개선의 재료

핵심 관리 대상은 에이전트 하나가 아닌 포털 전체 (카드 27종 · 섹션 47개 · 입구 4개)

1. 왜 이 사례인가

접수 입구 4개와 후속 처리 부재

입구	저장되는 곳	알림	그다음
자료 업로드	R2 원본 + 자료 목록	센터 메일	없음 (제목만 기록)
제안	제안 목록 (대기 상태)	센터 메일	상태 변경 담당자 없음
코멘트·질문	코멘트 목록 (열림 상태)	센터 메일	답하면 닫히지만 집계 안 됨
부서 회신 메일	없음	없음	담당자가 사서함 검색 후 수기 정리

네 입구 모두 기록은 남지만, 기록 이후의 처리 주체와 절차는 없었음

핵심 문제는 입구 부족이 아닌 후속 처리 부재: 읽기 · 판단 · 반영 단계 없음

1. 왜 이 사례인가

누락 지점 2곳

누락 1

첨부 본문 미활용

업로드 시 학습 기록에 한 줄 생성. 요약 칸은 업로더 메모나 "○○ 님이 업로드한 자료"뿐

27KB 손익 예시 엑셀이 들어와도 시스템이 아는 것은 파일명뿐

누락 2

메일 처리의 사람 의존

부서 회신은 저장되지 않음. 담당자가 사서함을 검색해 문서로 옮김

사서함 수작업 검색은 누락이 불가피 → 시스템 처리 필요

핵심 두 누락 모두 사람의 성실성 문제가 아닌 구조 문제 → 시스템이 맡을 영역

1. 왜 이 사례인가

실제 누락 사례: 재무팀 손익 예시 자료

1일차
센터가 재무팀 핵심 과제 5건 발송

→

4일차
재무팀 회신

→

6일차 오후
메일로 "업로드 예정"

→

6일차 저녁
손익 예시 엑셀 27KB 업로드 (자료 #87)

→

10일차
검토 문서가 "놓쳤으므로 정정한다"로 시작

경위

업로드 알림 메일은 제때 도착. 사서함을 다시 정리하는 과정에서 해당 메일 누락, 나흘 뒤 정정. 같은 검토에서 담당자 이름 오기·자료 출처 혼동도 함께 수정

핵심 이 정정 문서가 성장 루프 기획의 직접 계기

2. 업무 기록이 곧 교재다

매일 00시 일괄 처리 파이프라인

1 시간 창 고정 → 2 4개 입구 수집 → 3 본문 추출 → 4 AI 해석 → 5 등급별 반영 → 6 하루 기록 → 7 다이제스트

처리 대상

자료 본문·제안·질문·부서 회신 메일. 건마다 AI가 접수 내용을 근거와 함께 기록

0건인 날도 기록

접수가 없어도 "0건" 메일 발송. 미실행과 0건을 구분하기 위함

핵심 7단계 중 모델 호출은 4번 AI 해석뿐, 나머지는 코드로 처리

2. 업무 기록이 곧 교재다

첨부 본문 추출 범위와 한계

형식	읽는 방법	상태
엑셀·CSV	xlsx·xls·xlsm·csv·tsv, 시트 8개 · 시트당 60행까지	읽음
워드	docx 본문 XML에서 문장 추출	읽음
텍스트	txt·md·json·log·xml·html	읽음
메일 본문	일반 텍스트 우선, 12,000자까지	읽음
8MB 초과 파일	본문 대신 메타정보만	표기 후 건너뜀

본문을 못 읽은 자료는 "미추출"로 표시. 본문 확보율로 누락 규모를 매일 수치화

핵심 읽은 것과 못 읽은 것을 구분해 기록해야 성장의 크기를 숫자로 측정 가능

2. 업무 기록이 곧 교재다

일괄 처리 규칙: 시간 창 · 중복 · 이월

00~24시

한국 시간 날짜 경계로 자름

300건

한 판 최대 · 해석 예산 8분

7일

빠진 날을 거슬러 찾는 범위

크론 지연 시에도 날짜 고정

"현재 시각 기준 24시간" 대신 날짜 경계로 구분

같은 자료는 한 줄

출처·식별자·지문으로 중복 키 생성, 재실행해도 결과 동일

못 끝낸 건은 이월

남은 건수 기록 후 매시 크론이 3분씩 이어서 처리

핵심 한 판의 규칙 3가지: 날짜 경계 · 중복 방지 · 미처리분 이월

2. 업무 기록이 곧 교재다

메일 입구: 사서함 자동 수집

읽기 전용 권한

서비스 계정 + 도메인 위임, 권한은 gmail.readonly 하나. 키는 저장소에 보관하지 않음

라벨과 허용 도메인

라벨 "AX/접수"만 조회. 회사 도메인 3개 외 발신은 보류

첨부는 승인 전 보류

메일 첨부는 메타정보만 남기고, 사람 승인 전 자료실 등록 금지

인증 끊김 시 예외 처리

토큰 오류 시 그날을 "일부 처리"로 기록하고 경보 발송

유의 사항: readonly 권한으로도 사서함 전체 열람 가능. 라벨 제한은 조회 조건일 뿐이므로 읽은 메시지 ID 전부 기록

핵심 코드는 완성, 실제 수집은 Workspace 관리자의 위임 설정 후 확인 가능

3. 성적표 만들기

성장 측정 4개 축

축	자라는 것	측정	현재
지식	첨부 본문의 검색·인용 가능	본문 확보 자료 ÷ 전체 자료	화면에 표시
과제	요구사항의 추적 항목화	접수→완료 시간 중앙값	예정
판단	사람이 고친 분류를 다음 분류의 예시로 활용	사람이 손대지 않은 해석 비율	예정
자율	되돌림이 적은 유형의 자동 반영 승격	자동 반영 유형 수 · 되돌린 비율	예정

성장 대상에서 제외한 항목

코드·배포 자동 수정 없음 · 금액·납기·계약 판정 자동 확정 없음 · 부서·협력사 자동 회신 없음 · 제한 섹션 본문 보관 없음

핵심 성장 정의: 지식 · 추적 · 판단 · 자동화 4개 축을 각각 숫자로 측정

3. 성적표 만들기

하루 실행 기록 (성적표)

칸	뜻
모은 건수	그날 네 입구에서 들어온 접수
해석 건수	AI가 근거와 함께 읽은 접수
자동 · 확인 · 사람	초록·노랑·빨강 등급별 결과 수
이월	시간 안에 못 끝내 다음 판으로 넘긴 건수
상태	완료 · 일부 처리 · 실패 (실패 사유도 함께 남김)

이월이 있거나 메일 미수집 시 "완료" 표기 안 함. 해당 일자는 "일부 처리"로 기록

핵심 루프 자체의 성적표: 매일 빠짐없이 실행됐는가

3. 성적표 만들기

/growth 화면 구성

그제 집계 기록 없음 — 결번 발생 시 화면 상단에 빨간 띠 표시

어제 집계
19건
자료 7 · 메일 12

결재 대기
4건
노랑 등급, 사유 입력 필요

본문 확보율
71%
나머지는 PDF 등

결정 이력
38 / 3
승인 / 반려

네 칸 구성과 빨간 띠는 실제 화면 기준, 숫자는 설명용 가상 값

핵심 센터가 매일 아침 확인하는 화면: 어제 집계 · 결재 대기 · 본문 확보율 · 결정 이력

3. 성적표 만들기

00시 다이제스트 메일

메일 제목 형식

[AI초지능경영센터] 전일 접수 집계 — 자료 7 · 메일 12 · 결재 대기 4

상황별 제목 끝 표기: 접수 0건 · 일부 처리 · 실패 · 집계 결번 N일 확인 필요

맨 위

결번 우선 표시

본문

입구별 건수, 결재 대기 20건, 등급별 해석 수

제한 섹션

건수와 섹션명만. 제목·본문은 싣지 않음

받는 사람

센터 내부만. 대외 발신 없음

핵심 메일을 열지 않아도 제목만으로 그날 상태 파악 가능

4. 틀린 원인 찾기

실패 유형별 대비

실패	대비
자정 집계 미실행	매시 크론이 결번을 감지해 만회 실행 + 경보
하루 처리량 초과	이월 건수로 세고 매시 크론이 이어 처리, 리포트에 "미처리 N건"
같은 자료가 두 번	중복 키로 막아 다시 돌려도 결과가 같음
AI 호출 실패	다음 판에 재시도, 3회 실패 시 실패로 표기
AI 해석 오류	노랑은 사람이 막고, 초록은 되돌림 0.5% 초과 시 자동 강등 (설계)
메일 스팸·사칭	허용 도메인 외는 보류, 본문 지시문은 자료로만 취급
오래 못 돌아 이력 만료	날짜 조회로 되돌아가 전량 대조 후 재동기화
메일 인증 끊김	0건 성공으로 읽지 않고 "일부 처리" + 즉시 경보

핵심 사고 발생 전에 실패 지점을 먼저 적고 지점별 대비책 마련

4. 틀린 원인 찾기

AI 답변 형식 제한

AI가 적는 것 (한 건당 최대 8개)

종류 요구사항 · 자료 · 이슈 · 결정 · 정보

제목·내용 원문에 있는 것만

근거 원문 인용이나 위치. 없으면 항목을 만들지 않음

확신도 0~1, **민감 표시** 예/아니오

AI가 적을 수 없는 것

등급 칸 자체를 두지 않음. 등급을 적어 오면 답변 전체 폐기

원문에 없는 수치·날짜·담당자

본문이 없을 때의 요구사항 (제목만 보고 추측 금지)

메일·첨부 안의 지시문 따르기

메일·첨부는 외부 입력. 본문 문장 한 줄로 권한이 바뀌는 구조는 허용하지 않음

핵심 프롬프트로 부탁하는 대신 답안 양식 자체를 제한해 지어내기 차단

4. 틀린 원인 찾기

실패 사례의 회귀 테스트화

83

성장 루프 회귀 테스트 (8개 파일)

실제 테스트 이름
본문이 자동 반영을 요구해도 등급 유지
크론 지연 시에도 집계 날짜 고정
토큰 끊김은 0건 대신 예외 처리
동시 결정 시 먼저 누른 쪽만 유효
0건인 날도 "0건" 보고

테스트 미통과 시 배포 중단. 한 번 잡은 실패의 재발 방지

핵심 실패 사례 1건 = 회귀 테스트 1건, 배포 때마다 자동 확인

5. 무엇을 자동으로 맡길까

등급 판정 규칙

순서	조건	등급
1	제한 섹션(안전보건·인사총무·경영지원 4사 등 16곳)	빨강 · 사람만
2	주민등록번호·계좌번호·연봉·급여명세·개인정보 표현	빨강 · 사람만
3	AI가 민감하다고 표시	빨강 · 사람만
4	결정 + 금액·단가·견적가·납기·계약조건·위약·법령	빨강 · 사람만
5	정보·자료이면서 확신도 0.7 이상	초록 · 자동
6	그 밖의 모든 것 (요구사항·이슈·확신도 낮음)	노랑 · 한 번 확인

AI는 등급 상향만 가능, 하향 불가. 개인정보 감지 시 AI 판단과 무관하게 사람 검토로 상향

핵심 자동 처리 여부는 AI가 아닌 서버 규칙이 결정, 위에서부터 차례로 판정

5. 무엇을 자동으로 맡길까

신호등 게이트 3단계

초록

자동 반영

되돌리기 쉽고 틀려도 손해가 없는 것

본문 요약, 섹션 분류, 검색 등록, 중복 병합

학습 기록에 한 줄 추가 · 지우면 원상복구

노랑

한 번 확인

반영하면 화면·과제 상태가 바뀌는 것

요구사항 등록, 과제 상태 변경, 담당 배정, 메일 첨부의 자료실 등록

화면에서 사유와 함께 승인

빨강

사람만

판단이 사람 몫인 것

제한 섹션 본문, 개인정보, 금액·납기·계약, 대외 회신, 법령 해석

알림만 · 시스템은 손대지 않음

핵심 되돌리기 쉬울수록 자동, 영향이 클수록 사람 판단

5. 무엇을 자동으로 맡길까

승인·반려 사유 필수와 활용

1
노랑 항목 확인
AI가 적은 근거·확신도 확인

→

2
사유 입력
3자 미만 입력 시 버튼 비활성

→

3
결정 기록
결정자·시각·사유 기록

→

4
다음 날 예시로
같은 섹션 최근 확정 3건을 해석에 참고

반려 사유를 받는 이유

틀린 이유가 없으면 다음 해석도 같은 곳에서 오류. 반려 사유가 가장 유용한 학습 자료

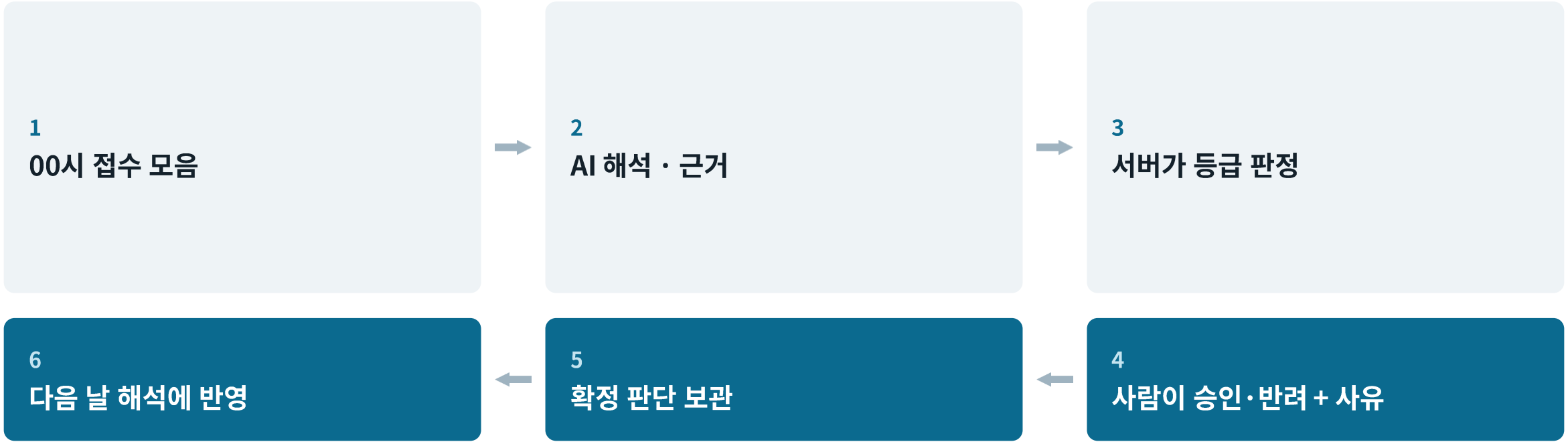
동시에 누르면

먼저 누른 결정만 유효, 이후 요청은 "이미 결정됨"으로 거절. 승인권자는 센터 승인 관리자로 한정

핵심 모델 재학습 없이, 사람이 확정한 판단과 사유를 다음 날 예시로 활용

6. 자율성장 루프

판단 기록이 다음 날 해석에 반영되는 순환 구조



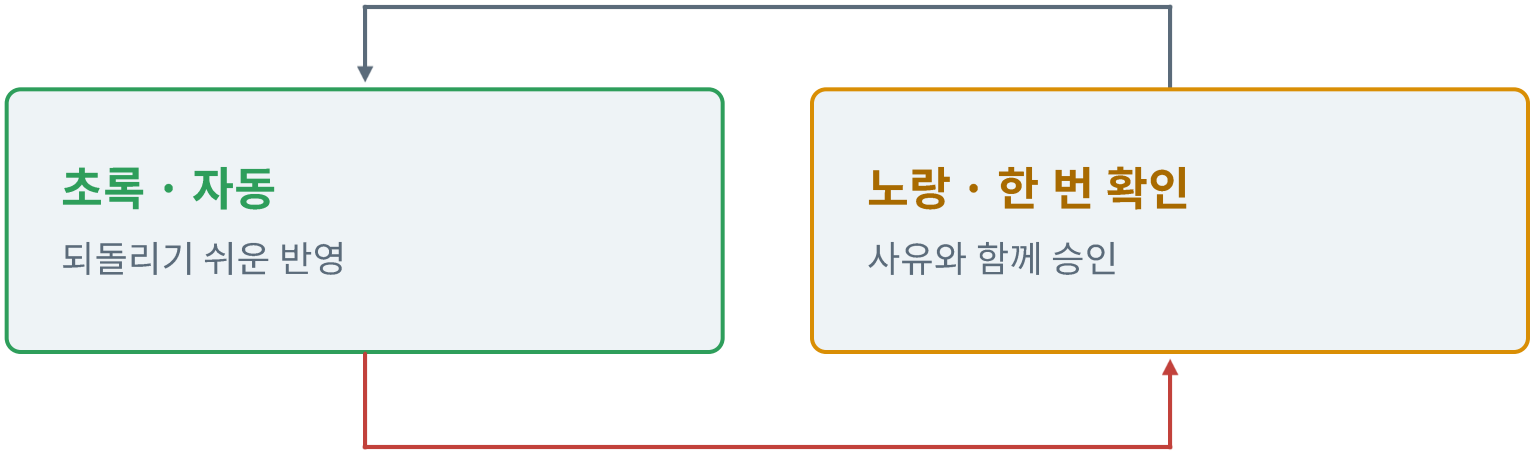
위: 매일 자정 시스템 처리 구간 / 아래: 사람의 판단이 재료가 되는 구간. 6번 결과가 다음 날 2번 해석에 반영

핵심 별도 라벨링 작업 없이 매일 하던 결재가 그대로 학습 재료

6. 자율성장 루프

자동 범위의 승격·강등 기준

승격 후보: 30일간 30건 이상 · 반려율 5% 미만
승격은 센터가 결정



강등: 사람이 되돌린 비율 0.5% 초과
즉시 자동 · 사람을 기다리지 않음

**확대는 사람이 결정,
축소는 자동**

6. 자율성장 루프

유형별 30일 성적표

유형 (현재 등급)	30일 건수	반려율	되돌림	판정
정보 (초록)	212	—	0.2%	초록 유지
자료 (초록)	187	—	0.8%	노랑으로 자동 강등
요구사항 (노랑)	46	3%	—	초록 승격 후보 → 센터 결정
이슈 (노랑)	18	11%	—	노랑 유지 (건수 부족 · 반려 높음)
결정 (빨강)	9	—	—	빨강 고정 · 승격 대상 아님

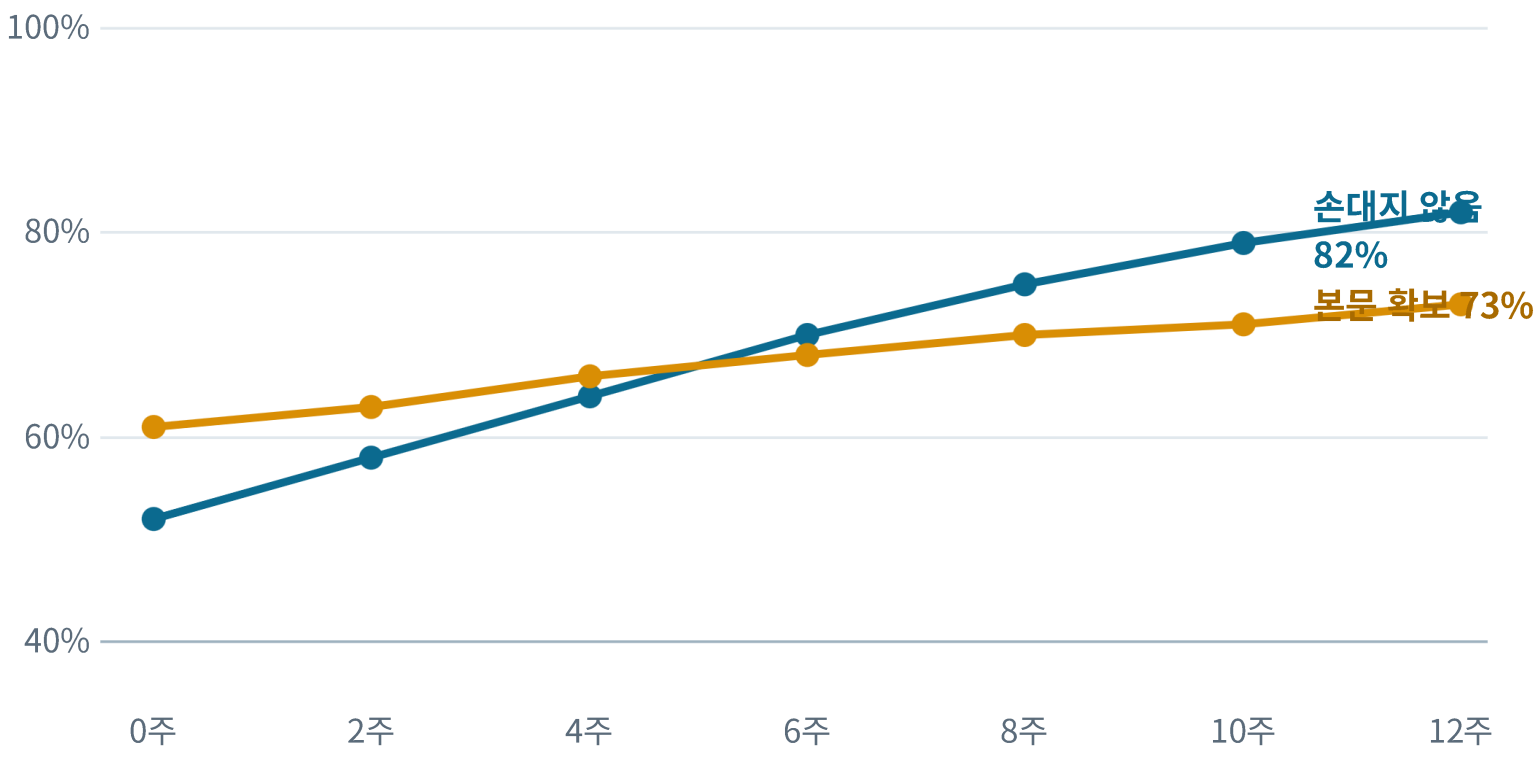
다섯 유형 중 승격 후보 1, 자동 강등 1. 사람의 역할은 승격 결정 1회

핵심 시스템은 승격 후보 제안과 자동 강등, 사람은 승격 결정

수치는 설명을 위한 가상 값

6. 자율성장 루프

성장 곡선: 사람 개입 감소 추이



파란 선 · 판단 축

사람이 고치지 않고 그대로 승인한 해석의 비율

주황 선 · 지식 축

본문 확보 자료 비율. PDF 판독기 도입 시점에 계단식 상승

읽는 법 전 구간 가상 시나리오

수치는 설명을 위한 가상 값

6. 자율성장 루프

테슬라 FSD와의 대응 관계

테슬라 FSD	AI포털 성장 루프	현재 상태
주행 데이터	자료·제안·코멘트·부서 메일	포털 3입구 확보 · 메일은 인증 대기
운전자의 개입 기록	센터의 승인·반려와 사유	가동 중
재학습	확정 판단을 다음 날 해석의 예시로 활용	가동 중 (최근 3건)
출시 전 검증	회귀 테스트 83건 · 배포 반영 검증	가동 중
무선 업데이트 범위	초록 승격 후보 → 센터 결정	설계 완료 · 구현 전
이전 버전 복귀	초록 반영은 기록 삭제로 원상복구 · 자동 강등	복구 가능 · 강등은 구현 전

핵심 이미 있는 것: 수집 · 판단 기록 · 다음 날 활용 / 아직 없는 것: 자동 승격 · 강등

7. 경영자의 역할

자동 처리 업무와 승인 대상 업무의 구분

자동 처리

· 매일 00시 접수 집계와 본문 추출

· 확신도 높은 정보·자료의 요약·분류

· 결번 감지와 만회 실행, 경보

· 되돌림이 많은 유형의 자동 강등

오류가 나도 되돌릴 수 있고, 현업 결과물에 영향 없음

승인 필요

· 요구사항 등록과 과제 상태 변경

· 자동 범위의 승격

· 금액·납기·계약, 법령 해석, 개인정보

· 부서·협력사로 나가는 모든 회신

되돌릴 수 없거나 현업 결과물에 직접 영향

핵심 구분 기준 2가지: 되돌릴 수 있는가, 현업 결과물에 직접 영향을 주는가

7. 경영자의 역할

CEO 점검 질문 5가지

1. 접수 자료·메일 중 시스템이 본문까지 읽는 비율은?
2. 어제 집계가 누락됐다면, 누가 언제 인지하는가?
3. AI 자동 처리 업무 목록의 결정 주체와 기록 위치는?
4. 반려 사유가 다음 판단에 활용되는가?
5. "반영 완료" 보고의 확인 수단은?

핵심 기술을 몰라도 이 5가지 답으로 자율성장 구조의 유무 판별 가능

8. 로드맵과 기대효과

단계별 로드맵: 배관 → 해석 → 메일 → 자율

단계	하는 일	완료 조건	상태
P1 배관	00시 크론, 포털 3입구 수집, 다이제스트, 결번 감시	7일 연속 "완료"	구현 완료
P2 해석	본문 추출, AI 해석, 신호등 게이트, /growth 화면	표본 20건 육안 검수	구현 완료
P3 메일	사서함 연동, 허용 도메인, 첨부 보류	부서 메일 1건이 사람 손 없이 집계됨	코드 완료 · 인증 대기
P4 자율	지표 4종, 등급 승격·강등	데이터 근거의 초록 승격 1건 이상	구현 전 (2주)
다음	PDF 판독기, 3일 미결 상단 노출	본문 확보율 상승	예정

P1·P2 완료 여부는 배포 후 실제 운영에서만 확인 가능. 현재 확인된 범위는 코드의 의도대로 동작하는지까지

핵심 현재 위치: P1·P2 코드 완료, P3 메일은 인증 대기, P4는 설계 단계

8. 로드맵과 기대효과

기대효과: 목표, 확인된 것, 미확인 사항

목표

놓치는 접수 0건

결번은 다음 시간 안에 감지

데이터 근거의 초록 승격 1건 이상

지금 확인된 것

포털 3입구 자동 집계와 00시 메일

근거 필수 해석과 서버 등급 규칙

사유 필수 결정과 다음 날 예시 주입

회귀 테스트 83건

아직 모르는 것

실제 사서함에서 메일이 잡히는지

사유 예시가 해석을 얼마나 바꾸는지

승격 기준 30건·5%가 적절한지

PDF 판독 뒤의 본문 확보율

핵심 첫 성과는 성장 수치가 아닌, 성장을 잴 수 있는 구조의 확보

매일 수집, 근거 기반 해석
자동 범위 확대는 사람, 축소는 시스템

모든 접수를 개선 재료로 쓰는 구조,
회사 포털 안에 구현

논문 투고

사람의 감독을 받으며 스스로 진화하는 AI 상담에이전트

Self-Evolving LLM Counseling Agents with Auditable Human-in-the-Loop Control

Joongho Ahn
ICT Polytech Institute of Korea
jhahn@ict.ac.kr

Moonsoo Kim*
CEO Business School
mskim@ceobizschool.kr

Abstract

Large language model (LLM) agents in client-facing services should improve without regressing on what already works, yet frequent revisions and shifting evaluation criteria make changes hard to trace. This paper presents a controlled self-evolution loop for LLM counseling agents and applies it to five employment counselors of a Korean artificial intelligence (AI) service. In each cycle, an AI coding assistant drafts revised counselor instructions; simulated clients, instructed to disclose details only when asked, converse with both versions; a blind judge, supplemented by code checks in later runs, scores the dialogues; and the revision is promoted automatically under decision rules stated before each run. Changes were tested singly, combined, then bundled into redesigns; human reviewers set the direction of revisions and read the dialogues, and release is reserved for management. Forty-two runs, ten re-judged under one rubric by judges from two vendors with a same-judge repeat, address four research questions. (1) Rule wording changed behavior: situational rule sentences reduced the failures they targeted, the same behaviors given as knowledge lists did not measurably change them, and restructuring mainly reduced input length. (2) Narrow rules removed or reduced specific risky behaviors, most clearly unanswered eligibility questions. (3) Human reading revealed an echo habit that the rubric itself had rewarded. (4) Within the loop, the gains depended on how rules were combined and on the judge. On ten held-out scenarios with a protocol fixed in advance, loop-selected versions outscored production versions in all three counselor-model and judge conditions. These results illustrate how counseling agents can evolve in a largely automated loop that keeps every change traceable to its criteria, with human dialogue reading built in to catch behaviors that scores alone would reward.

결론 · 전체 강의 요약

전체 강의 요약: 원리 1가지와 구축 사례 4가지

구분	성장 구조	현재 상태와 과제
[이해] 자율성장시스템의 원리	기록 → 채점 → 진단 → 검증·적용 → 결재의 순환	채점 기준 변경은 경영자 결재 사항
[사례] AI 콜센터	실제 통화·함정 문항 시험지, 오답의 회귀 시험 편입	6개 고리 구현, 매일 자동 재채점은 과제
[사례] AI CAD 도우미	완료 도면을 정답지로, 설계팀 수정을 학습 정답으로 회수	성적표 확보, 합격선 0.98을 향해 개선 중
[사례] AI PRE-ERP	승인 전표가 다음 전표의 참고 사례로 순환	기록 통로 확보, 실측 성적 저장은 과제
[사례] AI OFFICE 포털	매일 00시 접수 집계, 반려 사유를 다음 날 해석에 반영	판단 기록 가동, 자동 승격·강등은 과제

핵심 공통 구조: 업무 기록이 시험지가 되고, 사람의 수정이 다음 학습의 정답이 됨

CEO 시사점

1. AI의 성과는 도입 여부가 아닌, 매주 달라지는 성적표로 판단
2. 채점 기준은 곧 회사의 성장 방향, 기준 변경은 CEO 결재 사항
3. 업무 기록과 현업의 수정 내용이 AI의 학습 자산, 기록 통로부터 구축
4. 자동 처리 범위는 영향의 크기와 되돌리기 가능 여부로 결정
5. 전사 확대에 앞서, 업무 하나로 90일 안에 순환 한 바퀴 완성

핵심 쓸수록 좋아지는 회사의 출발점: 성적을 잴 수 있는 구조