CONDERI

ACADEMIC & PAPER RESEARCH INSIGHT | 대외 공개판

# 위임의 회계:
# 무엇을 AI에게 맡길 수 있는지는 모델이 아니라 검증 절차가 정한다

The Accounting of Delegation — Verification, Not Capability, Sets the Boundary

**이번 호의 문제의식**

지난 두 달의 실증 연구는 같은 결론으로 수렴합니다. AI가 할 수 있는 일의 범위를 넓히는 것은 모델 성능이 아니라 **사람이 어디까지 검토하고 무엇을 승인하는가**입니다. 이번 호는 네 영역 (기술·경영·과학·교육)에서 그 명제를 수치로 검증한 2026년 11월 공개 논문 7편을 다룹니다. 자동화율 2.5%→24.2%를 만든 것은 더 좋은 모델이 아니라 누적된 전문가 판단이었고, 조직 민첩성을 0.495만큼 끌어올린 자율 위임은 책임 공백도 0.342만큼 키웠습니다. 교실에서는 튜터 발화의 91.5%를 차지한 교사 중심 대화가 자동 피드백으로도 바뀌지 않았습니다.

**24.17%**
전문가 검토 누적 후
자동화율 (초기 2.50%)

**1.2~4.7배**
다중 에이전트 구조의
토큰·지연 비용 배수

**β 0.495**
자율 위임 → 조직 민첩성
(책임 공백 β 0.342)

**91.5%**
통제군 튜터의 발화 점유율
자동 피드백 후에도 동일

# 이번 호 논문 7편 한눈에

영역별 표본·방법·핵심 발견 — 숫자로 확인된 위임의 경계

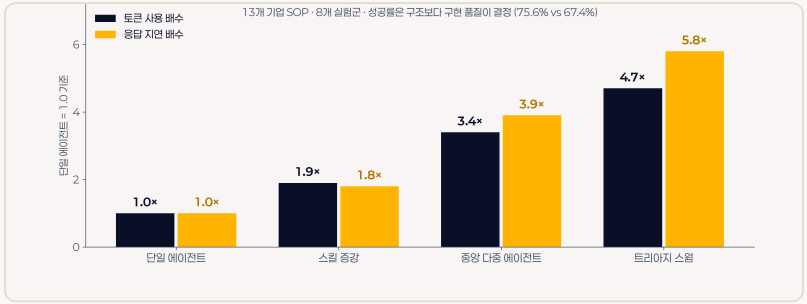
| 영역 | 논문 | 표본·방법 | 핵심 발견 |
|---|---|---|---|
| 기술 | Agent Architecture Cost–Success Trade-offs (arXiv 2611.12953, 2026-11-07) | 기업 SOP 13개 · 8개 실험군 · 모델·도구 고정 | 보편적 최적 구조 없음. 다중 에이전트는 토큰 1.4~4.7배·지연 2.6~5.8배, 성과 이득은 작고 유의하지 않음 |
| 기술 | 통계적 위임 프레임워크 (arXiv 2611.01566, 2026-11-01) | MultiHiertt 금융 QA · 생성–검증 다회 구조 | 전문가 판단 누적 시 자동화율 2.50%→24.17%, 릴리스 오류 1.20% 유지 |
| 경영 | Agentic AI as a Digital Employee (Zenodo 23862456, 2026-11-01) | 기업 리더 482명 · 4개 산업 · 24개월 종단 · PLS-SEM | 자율 위임 → 민첩성 β 0.495, 책임 공백 β 0.342. 진단적 통제는 효력 상실 |
| 경영 | 생성형 AI 협업과 의사결정의 질 (인문사회과학연구 34(1), 2026) | 국내 직장인 400명 · SmartPLS 4.0 | AI 협업은 의사결정의 질을 통해서만 팀 효율성에 도달(완전 매개). 노력기대는 유의하지 않음 |
| 경영 | 생성형 AI 활용 유형과 조직성과 (한국빅데이터학회지 11(1), 2026) | 생성형 AI 사용 경험 근로자 · 구조방정식 | 텍스트형은 정보처리·과부하 완화, 멀티모달형은 아이디어 창출에 강함. 둘 다 분석가용시간 경유 |
| 과학 | 조직적 난이도와 AI 벤치마크 (arXiv 2611.03494, 2026-11-02) | 조직이론 기반 분류체계 · 주요 벤치마크 전수 분석 | 주요 벤치마크 중 조직 차원을 주장한 비율 4분의 1 미만 |
| 교육 | 교실 발화 자동 피드백 RCT (arXiv 2611.07072, 2026-11-04) | 튜터 1,030명 · 학습자 7,443명 · 세션 6,508개 | 질문 유도·아이디어 수용·추론 8~15% 증가. 튜터 발화 점유율 91.5%는 불변, 학습 성과 유의차 없음 |
| 교육 | AI 튜터 확장 RCT (Google LearnLM·Eedi, 2026-11) | 영국 10개교 1,179명 · 4개 군 · 10주 | 근전이·중간 퀴즈 개선, 표준화 검사는 +0.02 SD로 유의차 없음. 초안 94.7% 무수정 승인 |

# I. 기술 — 에이전트 구조를 늘리는 일은 비용을 사는 일이다

SOP 13개 · 8개 실험군: 보편적 최적 구조 없음 · 다중 에이전트 토큰 1.4~4.7배 · 구현 품질이 결정

**문제의식: 구조를 늘리면 좋아지는가**

기업은 표준업무절차(SOP)를 실행하는 데 언어모델 에이전트를 씁니다. 그런데 어떤 구조가 나은지에 대한 통제된 증거는 거의 없었습니다. 이 연구는 모델·과제·도구를 고정하고 구조만 바꿨습니다. 함수호출 단일 에이전트, 스킬을 붙인 단일 에이전트, 중앙집중 다중 에이전트, 트리아지 스윔의 네 설계를 8개 실험군으로 나눠 기업 SOP 13개에 적용했습니다.



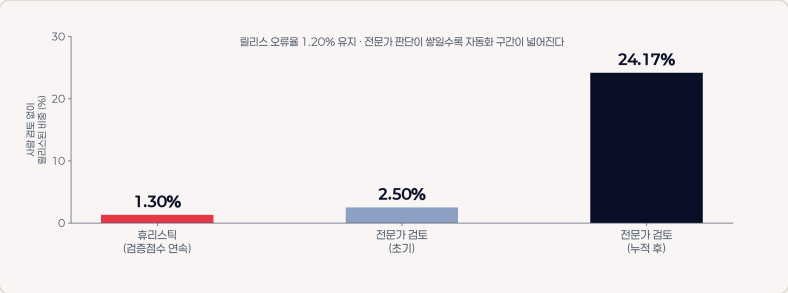

자료: Nandi (2026) arXiv 2611.12953. 배수는 단일 함수호출 에이전트를 1.0으로 둔 상대값 중간범위

**세 가지 결론**

**첫째, 보편적 최적 구조는 없습니다.** 같은 구조도 SOP에 따라 성공률이 10% 미만에서 100%에 가깝게 갈렸고, 전체 평균으로는 어떤 구조도 통계적으로 분리되지 않았습니다. 구조를 더한 것이 도움이 되는 것은 복잡하고 분기 많은 SOP에서만 방향성 있는 추세로 나타났습니다. **둘째, 구조의 비용은 성과로 정당화되어야 합니다.** 조정이 들어간 모든 실험군은 토큰을 1.4~4.7배, 지연을 2.6~5.8배 더 썼습니다. **셋째, 구현 품질이 구조 선택보다 중요합니다.** 평범하게 구현된 다중 에이전트는 단일 에이전트에 졌고, 같은 구조를 표적 재설계한 실험군만 기준선을 유의하게 이겼습니다(75.6% 대 67.4%). 스킬 정의의 품질도 성과를 갈랐습니다. 고품질 스킬군은 67.4%에서 73.1%로 올랐지만, 품질이 낮은 스킬군은 기준선 아래로 떨어졌습니다.

# II. 기술 — 자동화율을 올린 것은 더 좋은 모델이 아니라 누적된 전문가 판단이었다

자동화율 2.50%→24.17% · 릴리스 오류 1.20% 유지 · 휴리스틱은 1.30%에서 멈춤




자료: Agarwal et al. (2026) arXiv 2611.01566 · MultiHiertt 금융 질의응답 사례

### 검토가 쌓이면 자동화가 열린다

이 연구는 사람이 무엇을 검토해야 자동화를 늘릴 수 있는지를 통계적으로 정의합니다. 구조는 단순합니다. AI가 답을 생성하고 검증기가 여러 번 확인·수정하는 파이프라인에서, 배포 전에 전문가가 일부 산출물을 평가합니다. 그 평가로 검증기 임계값과 재시도 횟수를 보정하고, 통계적으로 승인된 후보 중 자동화율이 가장 높은 정책을 고릅니다. 생성기나 검증기를 바꾸지 않고도 릴리스 오류율을 통제할 수 있다는 점이 핵심입니다.

### 숫자가 말하는 것

금융 질의응답 데이터에서 전문가 평가가 누적되면서 사람 검토 없이 릴리스된 비중은 2.50%에서 24.17%로 늘었고, 릴리스된 산출물의 오류율은 1.20% 수준을 유지했습니다. 반면 "검증 점수가 연속으로 높으면 내보내는" 휴리스틱은 5% 오류 허용선을 넘기거나, 넘기지 않으려면 자동화율이 1.30%를 넘지 못했습니다. 같은 모델, 같은 데이터에서 **절차의 설계가 자동화의 크기를 10배 가까이 바꾼** 것입니다. 논문은 목표 자동화율을 달성하는 데 필요한 전문가 평가량을 정량화하고, 그 한계가 최적임을 보이는 하한까지 제시합니다.

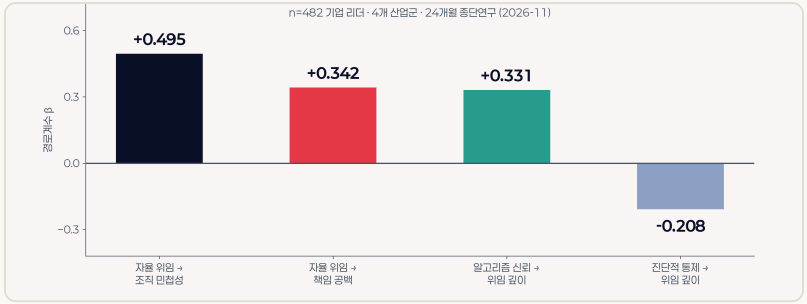
**핵심** **자동화율은 모델의 성질이 아니라 검토 예산의 함수다 — 검토를 늘리면 자동화가 늘어난다**

# III. 경영 — 자율 위임은 민첩성 0.495와 책임 공백 0.342를 함께 산다

리더 482명 · 4개 산업 · 24개월 종단연구 · PLS-SEM · 인터뷰 38건 · 다중 사례 4건

### 디지털 직원이라는 개념의 대가

4개 산업군(금융·헬스케어·공급망·클라우드/SaaS)의 리더 482명을 24개월간 추적한 종단연구입니다. 에이전트를 소프트웨어가 아니라 "디지털 직원"으로 규정하는 순간, 고전적 대리인이론과 관리통제 구조의 전제가 깨집니다. 연구진은 자율 위임이 조직 성과와 책임 구조에 동시에 작용한다는 것을 구조방정식으로 확인했습니다.




자료: Naseri et al. (2026) Zenodo 23862456 · PLS-SEM 경로계수 (모두 p<0.001 수준에서 유의)

### 민첩성의 대가

자율 위임은 조직 민첩성에 β 0.495로 유의한 정(+)의 영향을 주지만, 동시에 책임 공백과 조직 표류에 β 0.342의 정(+)의 영향을 줍니다. 성과와 책임이 같은 손잡이로 움직입니다. 위임을 늘리는 결정은 반드시 책임 구조의 재설계와 함께 가야 합니다.

### 통제의 실패

진단적 통제(결과 측정 후 시정)와 경계 통제(허용 범위 설정)는 비결정적 계획과 재귀적 API 호출 앞에서 효력을 잃습니다. 대신 현장 관리자는 과신(인지적 포기)과 미시 감시(알고리즘 냉소) 사이를 오갑니다. 두 반응 모두 처리량 이득을 상쇄합니다.

### 합성된 책임 구역

연구진은 중간관리자가 기술적 가시성도, 시간적 여유도 없이 발생한 에이전트 오류의 책임을 지는 구조를 "합성된 도덕적 완충지대"로 명명했습니다. 국내 조직에도 그대로 적용되는 명제입니다. 위임의 범위를 정할 때 책임의 착지점을 함께 정해야 합니다.

# IV. 경영 — 국내 실증 2편: 성과는 도구가 아니라 판단과 시간에서 나온다

직장인 400명 완전 매개 · 활용 유형별 경로 분해 · "AI를 쓴다"는 측정 단위가 아니다

### 국내 연구 1 — 의사결정의 질이 유일한 통로

국내 직장인 400명을 대상으로 한 연구(유정화·서영욱, 인문사회과학연구 34(1))는 생성형 AI 협업이 팀의 연대협력 업무효율성에 어떻게 도달하는지 검증했습니다. 결과는 명확합니다. AI 협업은 효율성에 직접 영향을 주지 않고, **의사결정의 질을 통해서만** 영향을 줍니다(완전 매개). 디지털 리터러시와 성과기대는 AI 협업을 유의하게 높였지만, 노력기대(쓰기 쉬울 것이라는 기대)는 유의하지 않았습니다. 쉬운 도구가 협업을 만들지 않고, 좋은 결정이 성과를 만든다는 뜻입니다.

### 국내 연구 2 — 활용 유형이 성과 경로를 가른다

한국빅데이터학회지 11(1)에 실린 연구(조영은·우종필·김중안)는 생성형 AI 활용을 텍스트 기반과 멀티모달 기반으로 나눠 경로를 분해했습니다. 텍스트 기반 활용은 정보처리역량 향상과 인지된 데이터 과부하 감소에 유의했고, 멀티모달 활용은 아이디어 창출역량에 더 큰 영향을 줬습니다. 두 유형 모두 분석가용시간을 늘렸고, 이 요인들이 비즈니스 분석역량을 거쳐 조직성과로 이어졌습니다. "AI를 쓴다"는 진술은 측정 단위가 아니라는 것이 두 연구의 공통 결론입니다.

| 구분 | 텍스트 기반 활용 | 멀티모달 활용 | 실무 번역 |
|---|---|---|---|
| **주 경로** | 정보처리역량 · 과부하 감소 | 아이디어 창출역량 | 문서·요약 조직과 시각·영상 조직은 다른 성과를 낸다 |
| **공통 경로** | 분석가용시간 증가 | 분석가용시간 증가 | 절약된 시간을 무엇에 쓰는지가 성과를 결정한다 |
| **최종 종속** | 비즈니스 분석역량 → 조직성과 | 비즈니스 분석역량 → 조직성과 | 도구 도입이 아니라 판단 역량이 종착점 |
| **시사점** | 데이터 과부하 조직에 우선 적용 | 기획·크리에이티브 조직에 우선 적용 | 부서별로 다른 도구 세트를 배치해야 한다 |

### 두 연구를 겹치면

AI 협업의 성과는 ①의사결정의 질 ②분석가용시간 ③활용 유형의 세 변수로 요약됩니다. 모두 **사람의 판단과 시간 배분**에 관한 변수이고, 모델 선택은 이 셋에 직접 등장하지 않습니다. 도입 예산의 배분을 도구 라이선스에서 판단 훈련과 검토 절차로 옮겨야 하는 이유입니다.

# V. 과학 — 조직 안의 일은 기술적 정답보다 어렵다: 벤치마크의 사각지대

조직적 난이도 5계층 분류 · 주요 벤치마크 중 조직 차원 주장 4분의 1 미만

### 벤치마크가 측정하지 않는 것

AI 벤치마크는 기술적 과제를 단위로 발전해 왔습니다. 그러나 경제적으로 가치 있는 일은 대부분 조직 안에서 일어납니다. 과제는 상호의존, 분산된 지식, 차등적 권한, 시간 제약, 상충하는 이해관계, 변하는 환경 속에 놓여 있습니다. 이 논문은 그 조건을 **조직적 난이도**라는 별도의 능력 차원으로 정의하고, 조직이론에 기반한 분류체계를 만듭니다.

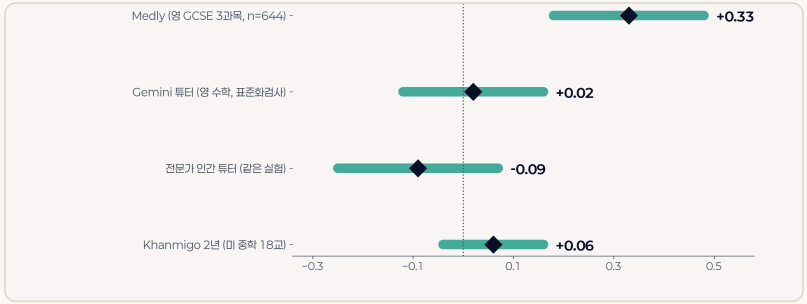
| 계층 | 구성 요소 | 에이전트가 넘어야 하는 것 |
|---|---|---|
| **작업 구조** | 상호의존성 · 분업 · 표준화 정도 | 내 과업이 다른 사람의 과업과 어떻게 연결되는지 인식 |
| **지식 구조** | 분산된 전문성 · 암묵지 | 누구에게 무엇을 물어야 하는지 판단 |
| **권한 구조** | 의사결정권 · 승인 · 에스컬레이션 | 결정권을 침범하지 않고 필요한 승인을 받는 절차 준수 |
| **거버넌스** | 조정 기제 · 규정 · 감사 | 규정과 감사 요구를 만족하는 방식으로 실행 |
| **성과** | 과업 품질 · 조정 성과 · 조직적 품위 · 자원·시간 | 정확한 답만으로는 부족하고, 절차와 시점이 함께 평가됨 |

### 절반의 결론, 절반의 경고

연구진은 주요 벤치마크를 분류체계로 분석해 조직 차원을 주장한 비율이 4분의 1에 못 미친다는 것을 확인했습니다. 기술적으로 정확한 산출물이 조직 안에서 성공을 의미하지 않는다는 것입니다. 권한을 우회하거나, 정보를 부적절하게 공개하거나, 의존 활동이 조정할 수 없을 만큼 늦게 행동하면 정답도 실패입니다. 이 구분은 **조직 향행 능력**과 **조직적 품위**로 나뉩니다. 능력이 있어도 규범을 어기면 성과가 아닙니다. 사내 에이전트 평가지를 만들 때 마지막 두 줄을 반드시 넣어야 하는 이유가 여기에 있습니다.

# VI. 교육 — 가까운 효과는 크고, 먼 효과는 아직 없다

튜터 1,030명·학습자 7,443명 RCT: 발화 유도 +8~15%, 발화 점유율 91.5% 불변 · 표준화 검사 +0.02 SD




자료: arXiv 2609.14789(Medly) · Google LearnLM·Eedi(2026) · NBER 중학 18교 연구 종합

### 교실에서 바뀐 것과 바뀌지 않은 것

SAT 수학 부트캠프 2회에 걸쳐 튜터 1,030명, 학습자 7,443명, 녹음 세션 6,508개를 무작위 배정한 실험입니다. 튜터에게 자동 피드백을 주자 학습자의 아이디어를 끌어내는 발화, 학습자 기여에 대한 수용, 학습자 추론이 8~15% 늘었고 두 캠프에서 같은 패턴이 재현됐습니다. 그러나 통제군에서 평균 91.5%였던 **튜터의 발화 점유율은 어떤 조건에서도 변하지 않았습니다.** 학습자에게 직접 준 참여 피드백도 대화 구조를 더 바꾸지 못했습니다. 통합 모형에서는 학습 성과의 유의한 차이가 없었습니다.

### 확장 실험: 표준화 검사 앞에서 사라지는 효과

영국 10개 학교 1,179명을 10주간 네 집단으로 나눈 사전등록 실험(Google LearnLM·Eedi)은 더 정교합니다. 튜터는 생성 초안 94,385건 중 94.7%를 수정 없이 승인했고, 77.1% 세션은 수정이 없었습니다. 감사한 114,935건 중 수학 오류·어조 문제는 6건(0.01% 미만)이었고 학생에게 도달하지 않았습니다. 오답 수정과 근전이는 전문가 튜터와 AI가 비슷하게 개선했습니다(57.7% 대 56.9%, 통상 43.4%). 그런데 외부 채점 표준화 검사에서는 어떤 집단도 통상 지원을 이기지 못했습니다. AI는 −0.01~+0.02 SD, 전문가 인간 튜터는 −0.09 SD였습니다.

# VII. 통합 — 위임의 네 칸: 과업 · 검증 · 승인 · 책임

네 영역의 논문을 하나의 도입 서식으로 — 30일 파일럿 설계

### 네 편을 하나의 설계로

네 영역의 논문은 서로 다른 대상을 보지만 같은 구조를 가리킵니다. 위임은 **과업** → **검증** → **승인** → **책임**의 네 칸으로 이루어지고, 자동화의 크기는 두 번째 칸(검증)의 설계로 결정됩니다. 아래 표는 사내 에이전트 도입 시 바로 채울 수 있는 서식입니다.

| 칸 | 적는 내용 | 논문이 주는 기준 |
|---|---|---|
| ① 과업 | SOP 단위로 분해한 과업, 분기 수, 필요한 도구 | 구조보다 SOP 적합성이 먼저 — 13개 SOP에서 성공률 10~100% |
| ② 검증 | 검증기 종류, 재시도 횟수, 표본 검토 비율 | 전문가 평가 누적으로 임계값 보정 — 자동화율 2.5→24.2% |
| ③ 승인 | 승인권자, 승인 없는 릴리스 금지 구간 | 릴리스 오류 1.20% 통제 · 휴리스틱은 1.30%에서 정체 |
| ④ 책임 | 오류 발생 시 책임의 착지점과 보고 시점 | 책임 공백 β 0.342 · 중간관리자 완충지대 방지 |

### 30일 파일럿 설계

첫 30일은 성과가 아니라 **측정 가능성**을 검증합니다. ① 과업 3개를 골라 SOP를 문장으로 씁니다. ② 각 과업에 검토 표본 비율을 10%로 두고 오류율을 기록합니다. ③ 2주 뒤 검증기를 고정한 채 검토 표본만 늘려 자동화율이 오르는지 봅니다. ④ 4주 뒤 승인 없는 릴리스 구간을 하나 정합니다. 이 순서를 지키면 모델을 바꾸지 않고도 위임 범위가 넓어지는지 확인할 수 있습니다.

**설계 원칙** 좋은 에이전트는 좋은 검토 절차에서 나온다 — 네 칸을 채우면 도구 선택은 마지막 문제가 된다

# English Summary / 영문 요약

Verification design, not capability, sets the boundary of delegation


## English Summary

Seven papers published in November 2026 converge on one proposition: the boundary of what can be delegated to AI is set by verification design rather than by model capability. Nandi (arXiv 2611.12953) holds model, tasks and tools fixed across thirteen enterprise standard operating procedures and eight architectural arms, and finds no universally best architecture: coordinated designs spend 1.4 to 4.7 times the tokens and 2.6 to 5.8 times the latency of a single agent for gains that are rarely significant, while implementation quality dominates the choice - a targeted redesign reached 75.6% task success against a 67.4% baseline. Agarwal and colleagues (arXiv 2611.01566) show that a statistical delegation policy calibrated by accumulated expert evaluations lifts the automation rate from 2.50% to 24.17% while holding released-output error near 1.20%, whereas score-threshold heuristics either breach a 5% error tolerance or automate at most 1.30% of tasks.

## Management, science and education

A 24-month longitudinal study of 482 enterprise leaders (Zenodo 23862456, PLS-SEM) finds autonomous delegation raises organisational agility (beta 0.495) and accountability drift (beta 0.342) at once, and that diagnostic and boundary controls degrade against non-deterministic agent planning, producing what the authors call a synthetic moral crumple zone around middle managers. Two Korean studies add mechanism: among 400 employees, generative AI collaboration improved team efficiency only through decision quality (full mediation), and a decomposition by usage type showed text-based use raising information processing while multimodal use favoured idea generation, both via analyst time. Barres and colleagues (arXiv 2611.03494) argue that organisational task difficulty is a separate capability dimension that fewer than a quarter of prominent benchmarks address.

## Classroom evidence

In a randomised trial across two SAT mathematics bootcamps (1,030 tutors, 7,443 learners, 6,508 sessions; arXiv 2611.07072), automated feedback increased eliciting of learner ideas, uptake and learner reasoning by 8 to 15%, yet tutor talk time - 91.5% of the control condition - did not change in any arm, and pooled learner outcomes showed no significant effect. An expanded UK trial of 1,179 students found near-transfer gains and a midtrial quiz advantage, but no arm beat standard remediation on an externally scored standardised test (-0.01 to +0.02 standard deviations for AI, -0.09 for expert tutors).


**One line** **Automation is a function of review budget, not of model quality - fill the four boxes before choosing a tool**

**발행** CONDERI AI산업전략연구실 | 문서등록번호 CONDERI-AR-2026-11-16 | 문의 contact@conderi.com