CONDERI

ACADEMIC & PAPER RESEARCH INSIGHT | 대외 공개판

성과를 가르는 것은 모델이 아니라 맥락과 증거다

Context and Evidence, Not Model Size — Four Fields, One Boundary

이번 호의 문제의식

같은 모델을 써도 성과가 갈리는 이유를 이번 달 논문들은 세 가지 숫자로 설명합니다. 첫째, 맥락입니다. 기업 데이터 분석 에이전트의 성과는 과업 유형에 따라 어떤 메타데이터를 주는지로 갈렸고, 사후에 필요 맥락을 채워 넣는 비율은 34.9%에 그쳤습니다. 둘째, 상호작용의 순서입니다. AI에게 먼저 묻는 워크플로는 AI를 제거한 뒤 추론 품질이 무AI 집단보다 낮았고(d=−0.28), 자기 입장을 먼저 정하는 워크플로는 그 하락이 없었습니다. 셋째, 증거의 단위입니다. 교실 무작위 실험의 표본이 1,300명을 넘어서면서, 실험실에서 나온 낙관적 효과가 표준화 검사에서는 사라지는 사실이 드러났습니다.

34.9%
사후 필요 맥락 포함률
에이전트 분석 실험

d −0.28
의존형 워크플로의
비보조 추론 하락

0.31 SD
나이지리아 튜터링
1,328명 무작위 실험

94.7%
전문가가 수정 없이
승인한 AI 튜터 발화

이번 호의 지도 — 네 분야, 하나의 질문

맥락 · 상호작용 순서 · 증거의 단위, 세 숫자가 위임의 경계를 정한다

네 분야, 하나의 질문

이번 호의 논문은 서로 다른 학회지에 실렸지만 같은 질문을 다룹니다. **AI가 잘하는 일과 사람이 책임질 일의 경계는 어디인가.** 기술 분야에서는 에이전트에 무엇을 알려주어야 하는지, 경영 분야에서는 어떤 조직 조건에서 AI 스킬이 성과로 전환되는지, 과학 분야에서는 AI 사용이 사용자의 추론을 어떻게 바꾸는지, 교육 분야에서는 교실에서 실제로 얼마나 배우는지를 측정했습니다. 네 답이 같은 방향을 가리킵니다. **성과는 모델의 능력이 아니라 입력의 구조와 검증의 절차에서 결정됩니다.**

분야	수록 저널·학회	이번 호에서 다루는 핵심 숫자	실무 질문
기술	지능정보연구 32(3) · 경영정보학회 추계	맥락 조합 8종 · LLM 3종 × 5회 반복	에이전트에 무엇을 먼저 줄 것인가
경영	경영학연구 55(5) · 55(4)	AI 스킬 → 사업화역량 조절 구조	어떤 조직에서 AI 스킬이 성과가 되는가
과학	Frontiers in Psychology · HSSC	3군 900명 · 구조방정식 698명	AI 사용이 추론을 어떻게 바꾸는가
교육	NBER · 영국·나이지리아 RCT	1,328명 · 1,179명 · 0.06~0.31 SD	교실에서 실제로 배우는가

이번 호를 읽는 순서

기술 두 장은 에이전트의 입력 설계를 다룹니다. 첫 장은 데이터 맥락이 분석 성과를 어떻게 가르는지, 둘째 장은 에이전트를 몇 개로 나눌 때 비용과 정확도가 어떻게 바뀌는지입니다. 경영 두 장은 조직 조건과 윤리 설계를, 과학 한 장은 인지 위임의 두 경로를 다룹니다. 교육 두 장은 교실 실험 네 건을 효과크기와 표본 크기로 비교합니다. 마지막 장에서 네 분야의 발견을 하나의 적용 프레임워크로 묶습니다.

논문	저자	수록	이번 호의 위치
기업 LLM 분석 에이전트의 데이터 맥락	박기준·조미영	지능정보연구 32(3) 449-473	I장 맥락 설계
MCP·RAG 계층형 멀티 에이전트	임송재	지능정보연구 32(3) 309-330	II장 추적성
싱글 vs 멀티 에이전트 금융보고서	박호연	지능정보연구 32(3) 39-56	II장 역할 분리
AI 스킬 → 기술사업화역량	김현정·김재명	경영학연구 55(5) 2011-2045	III장 조직 조건
AI는 관리자 윤리를 어떻게 재구성하는가	김성준·이중학	경영학연구 55(4) 1517-1556	IV장 기본값
Borrowed or kept?	Liu 외	Frontiers in Psychology	V장 인지 위임
One Click Away: Khanmigo 2년	Oreopoulos·Low	NBER w35620	VI장 교실 실험

I. 기술 — 에이전트의 병목은 모델이 아니라 맥락이다

지능정보연구 32(3): 과업 유형별로 필요한 맥락이 다르고, 사후 보충은 34.9%만 작동한다

같은 스키마, 다른 성과

지능정보연구 32권 3호에 실린 박기준·조미영의 연구는 기업 데이터 분석 에이전트의 실패 원인을 모델이 아니라 입력에서 찾습니다. 자연어-SQL 분석에는 테이블 스키마뿐 아니라 컬럼 의미, 업무 용어, 지표 산식, 조인·집계 규칙이 필요합니다. 연구는 Task-Technology Fit과 Cognitive Fit을 바탕으로 Context-Task Fit을 개념화하고, 통신 도메인 합성 데이터의 held-out 과업으로 통제 실험을 했습니다. Study 1은 확장 기술 메타데이터·비즈니스 글로서리·시맨틱 레이어의 8개 조합을, Study 2는 분석 가능한 네 조건을 세 개 LLM과 5회 반복으로 평가했습니다.

과업 유형	가장 효과가 큰 맥락	관찰된 패턴
단순 조회	기본 스키마	기본 스키마만으로 천장효과 — 맥락을 더해도 개선 없음
지표 계산	확장 기술 메타데이터 포함	시맨틱 레이어 단독보다 높은 성과
업무 해석	시맨틱 레이어	업무 용어·지표 정의가 핵심 동인
글로서리 단독	효과 제한	실행 규칙·컬럼 연결을 배제한 자연어 정의만으로는 부족

맥락은 많이 주는 것이 아니라 맞게 주는 것

가장 실무적인 숫자는 34.9%입니다. 사전에 매핑하지 않고 사후에 필요한 맥락을 채워 넣은 비율이 34.9%에 그쳤습니다. 과업유형 기반 사전 매핑도 전체 제공보다 낮았습니다. 즉 "일단 다 주고 모델이 고르게 하는" 방식은 작동하지 않았습니다. 반대로 사전 매핑 외에 맥락을 추가했을 때 일관된 성능 악화는 나타나지 않았습니다. 오류 정보에 의존하는 과업에서만 상대적으로 큰 정확도 하락이 관찰됐습니다. 결론은 명확합니다. **맥락은 데이터 카탈로그의 문제가 아니라 과업 정의의 문제입니다.**

데이터 카탈로그를 다시 쓰는 법

실무에서 바로 할 수 있는 것은 세 가지입니다. ① 지표 정의서에 분자·분모·필터·조인 규칙을 명시합니다. 시맨틱 레이어가 업무 해석 과업의 핵심 동인이었다는 결과는 이 문서가 곧 성능이라는 뜻입니다. ② 과업 유형별로 필요한 메타데이터를 매핑한 표를 만들어 프롬프트에 고정합니다. 단순 조회에 계산용 맥락을 붙이면 토큰만 늘고 성과는 그대로입니다. ③ 새 질문이 실패하면 스키마 오해·용어 불일치·조인 오류·해석 오류 네 가지로 분류해 카탈로그에 반영합니다. 네 가지 분류가 쌓이면 다음 에이전트의 프롬프트가 저절로 좋아집니다.

II. 기술 — 멀티 에이전트의 값은 정확도가 아니라 추적성

경영정보학회 추계 120여 편과 지능정보연구가 같은 곳을 가리킨다 — 검증 가능성

에이전트를 나누면 무엇이 좋아지는가

같은 호에 실린 박호연의 연구는 싱글 에이전트와 멀티 에이전트로 금융투자보고서를 생성해 성능을 비교했습니다. 임송재의 연구는 재난문자 발송 관리자의 의사결정을 돕기 위해 MCP와 RAG를 결합한 계층형 멀티 에이전트 시스템을 설계했습니다. 두 연구가 공통으로 보여주는 것은 역할 분리의 효과가 정확도보다 **검증 가능성**에서 크다는 점입니다. 생성·검증·요약을 나누면 어느 단계에서 오류가 생겼는지 추적할 수 있고, 그 추적이 사람의 승인 판단을 빠르게 만듭니다.

학회의 주제가 옮겨간 곳

한국경영정보학회는 11월 12~13일 연세대에서 추계학술대회를 열었습니다. 주제는 "Agentic AI에서 Physical AI로"였고 논문 120여 편이 발표됐습니다. AI 전환과 업무 혁신, 생성형 AI와 조직 변화, 인간과 AI의 협업이 주요 트랙이었습니다. 별도 세션에서는 AI 코드 어시스턴트의 공급망 보안, AI 안경 착용 시 주변인의 프라이버시, AI 시스템의 신뢰성 검증이 다뤄졌습니다. 에이전트의 성능이 아니라 **에이전트가 남기는 흔적**이 학회의 관심사로 옮겨간 것입니다.

120여 편
경영정보학회 추계 발표 논문

32(3)
지능정보연구 최근 호 · 28편 수록

3계층
MCP·RAG 기반 재난문자 의사결정 구조

2종
싱글 vs 멀티 에이전트 보고서 생성 비교

운영 비용은 어디서 나오는가

멀티 에이전트는 토큰을 더 쓰고 지연을 늘립니다. 그 비용을 정당화하는 것은 자동화율이 아니라 검토 시간의 감소입니다. 역할을 나눈 구조에서는 각 단계의 입출력이 로그로 남고, 검토자는 전체를 다시 읽지 않고 실패 지점만 확인할 수 있습니다. 반대로 역할이 하나인 구조는 빨리 끝나지만, 틀렸을 때 어디가 틀렸는지 알려주지 않습니다. 조직이 실제로 사는 것은 속도가 아니라 책임의 위치입니다.

에이전트 수를 정하는 세 가지 기준

첫째, 검증 지점이 세 개를 넘는가. 생성·검증·요약이 각각 다른 사람의 승인을 받아야 한다면 역할을 나눌 이유가 있습니다. 둘째, 실패 비용이 단계별로 다른가. 초안 오류와 최종 승인 오류의 비용이 다르면 단계별 로그가 필요합니다. 셋째, 로그를 사람이 읽을 수 있는가. 읽을 수 없는 로그는 추적성이 아니라 저장 비용입니다. 세 기준에 해당하지 않는데 에이전트를 나누면 토큰과 지연만 늘어납니다.

III. 경영 — AI 스킬의 가격은 조직이 정한다

경영학연구 55(5): 전략적 유연성과 위험관리가 스킬과 성과 사이를 조절한다

스킬이 역량이 되는 조건

경영학연구 55권 5호에 실린 김현정·김재명의 연구는 구성원의 AI 스킬이 기술사업화역량으로 전환되는 조건을 다룹니다. 제목이 밝히는 구조는 두 가지입니다. 전략적 유연성이 그 전환의 강도를 조절하고, 위험관리가 조절된 매개 경로에 개입합니다. 즉 AI 스킬을 많이 가진 조직이 성과를 내는 것이 아니라, **전략을 바꿀 수 있는 조직**에서 스킬이 사업화 역량으로 번역됩니다. 연구는 설문 기반 구조방정식으로 이 경로를 검증했습니다.

구성	연구가 검증한 관계	실무 해석
AI 스킬	기술사업화역량의 선행 요인	교육·채용만으로는 성과가 되지 않는다
전략적 유연성	스킬 → 역량 전환의 조절변수	사업 방향을 바꿀 권한이 있어야 스킬이 쓰인다
위험관리	조절된 매개 경로에 개입	실패를 흡수하는 장치가 있어야 실험이 가능하다
시사점	조직 조건이 스킬의 가격을 정한다	AI 교육 예산보다 의사결정 구조를 먼저 손본다

같은 호의 다른 두 논문

이대성은 유통·물류 산업에서 데이터 2차 사용이 조직 간 자원 보완성을 통해 이중 가치를 창출하는 경로를 단일 사례로 분석했습니다. 데이터를 보유한 쪽과 활용할 수 있는 쪽이 다를 때, 가치는 데이터 판매가 아니라 **마찰을 줄이는 정착 경로**에서 발생한다는 관점입니다. 정혁진·김한구는 AI 에이전트의 제품 추천에서 응답 지연이 수용 의도에 미치는 영향을 나르시시즘 성향에 따라 다르게 해석되는 시간적 단서로 분석했습니다. 같은 2초가 모든 사용자에게 같은 의미가 아니라는 것입니다. 세 논문을 함께 놓으면, AI 성과의 조건은 기술이 아니라 조직·데이터·사용자 성향의 삼각형 안에 있습니다.

조절
전략적 유연성 — 스킬→역량 전환의 강도

매개
위험관리 — 조절된 매개 경로에 개입

설계
설문 기반 구조방정식 검증

시사
AI 교육 예산보다 의사결정 구조를 먼저

IV. 경영 — 윤리와 지연의 설계

AI가 결정을 초안으로 바꾸면, 책임은 승인 버튼의 위치에서 갈린다

AI는 관리자의 윤리를 어떻게 재구성하는가

김성준·이종학의 연구(경영학연구 55권 4호)는 AI가 관리자의 윤리 판단 구조 자체를 바꾸는 방식을 다룹니다. 알고리즘이 후보를 정렬하고 평가 초안을 만들면, 관리자의 판단은 "결정"에서 "수정"으로 이동합니다. 문제는 수정이 기본값이 될 때 생깁니다. 초안을 그대로 승인하는 것이 가장 쉬운 선택이 되고, 책임은 초안을 만든 시스템과 승인한 사람 사이에서 흐려집니다. 논문은 이 구조 변화를 윤리적 판단의 재구성으로 설명합니다.

지연의 설계: 같은 2초가 다르게 읽힌다

정혁진·김한구의 연구는 AI 에이전트의 응답 지연이 수용 의도에 미치는 영향을 나르시시즘 성향에 따라 분석했습니다. 사람은 대기 시간을 "시스템이 나를 위해 생각하는 시간"으로 읽기도 하고 "시스템이 나를 우선하지 않는 신호"로 읽기도 합니다. 같은 지연이 어떤 사용자에게는 정성의 증거, 다른 사용자에게는 무시의 증거가 됩니다. 인터페이스의 시간 설계가 곧 평가의 설계라는 뜻입니다.

설계 변수	관찰된 문제	통제 방법
기본값 승인	초안 그대로 승인하는 관성	승인 전 최소 1개 항목 수정 요구, 수정 로그 보관
책임 공백	시스템과 승인자 사이의 책임 분산	판단 항목별 승인자 기록, 반려 사유 코드화
응답 지연	사용자 성향에 따른 해석 차이	대기 중 처리 상태 표시, 예상 시간 명시
설명 부재	점수·순위의 근거 미제시	상위 3개 근거와 반대 근거 함께 노출

윤리는 문서가 아니라 기본값이다

세 편을 종합하면 실무 원칙은 네 줄로 줄어듭니다. 기본값을 승인으로 두지 말 것, 판단 항목마다 사람의 이름을 남길 것, 대기 시간을 설명할 것, 점수의 근거를 함께 보여줄 것. 정책 문서로는 이 네 가지를 강제하기 어렵습니다. 제품의 기본 설정으로 넣어야 지켜집니다.

평가자가 사라진 자리

AI가 후보를 정렬하고 초안을 만들면, 평가의 기준은 화면 밖으로 밀려납니다. 문제는 그 기준이 문서로 남지 않는다는 점입니다. 11월 10일 한국경제인협회 조사에서 신입 채용에 AI 활용 역량을 공식 평가 항목으로 반영하는 기업은 20.0%였고, 전 직무의 필수 항목으로 두는 기업은 1% 수준이었습니다. 기준이 없으면 승인은 관성이 되고, 관성은 책임을 흐립니다. 윤리 논문이 말하는 재구성은 거창한 원칙이 아니라 평가표의 한 줄에서 시작합니다.

V. 과학 — 인지 위임의 두 경로

몰입은 사고를 돕고 의존은 사고를 대체한다 — 같은 도구, 다른 결과



자료: Liu 외(2026) Frontiers in Psychology 3군 실험(900명 배정·659명 분석) · Tian·Zhang(2026) HSSC 구조방정식(698명)

묻는 순서가 추론을 바꾼다

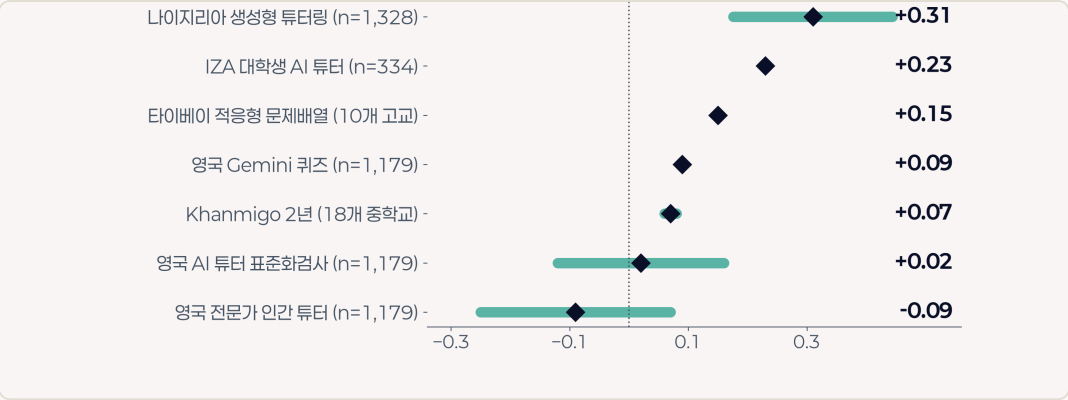
3군 실험(900명 배정, 659명 분석)은 AI 사용의 양이 아니라 **순서**를 조작했습니다. 자기 입장을 먼저 정하고 AI에게 질문하는 워크플로, AI 답변을 먼저 받아들이는 워크플로, AI를 쓰지 않는 세 집단입니다. 15~20분 뒤 AI를 모두 제거하고 같은 주제의 두 번째 글을 쓰게 했습니다. 의존형 집단은 AI를 한 번도 쓰지 않은 집단보다 추론 품질이 낮았습니다(1~7점 척도에서 −0.38, d=−0.28, p=0.0002). 자기생성 우선 집단은 기준선 아래로 떨어지지 않았습니다.

몰입은 이롭고 의존은 해롭다

중국 대학생 698명 구조방정식에서는 두 경로가 동시에 나타났습니다. 인지된 AI 지능은 집중 몰입을 높여 비판적 사고를 개선했습니다(β=0.50, 몰입 경유 0.25). 반면 몰입이 의존을 키우면 비판적 사고가 낮아졌습니다(의존 → 비판적 사고 β=−0.25, 순차 매개 −0.104). 193명 실험실 연구도 같은 방향입니다. AI 지원은 과제 정확도를 높였지만, 도구가 사라진 뒤에는 노력 조절이 둔해졌습니다. 세 연구의 공통 결론은 **사용량이 아니라 개입의 순서**입니다.

VI. 교육 — 교실 실험이 실험실을 이겼다

효과크기 0.06~0.31 SD, 그러나 표준화 검사에서는 사라진다 — 병목은 사용이다



자료: 각 논문 공개 초록·보고서. 막대는 공개된 신뢰구간, 점은 보고된 점추정치. 효과크기는 표준화 단위

근거접합과 표준화 검사의 간격

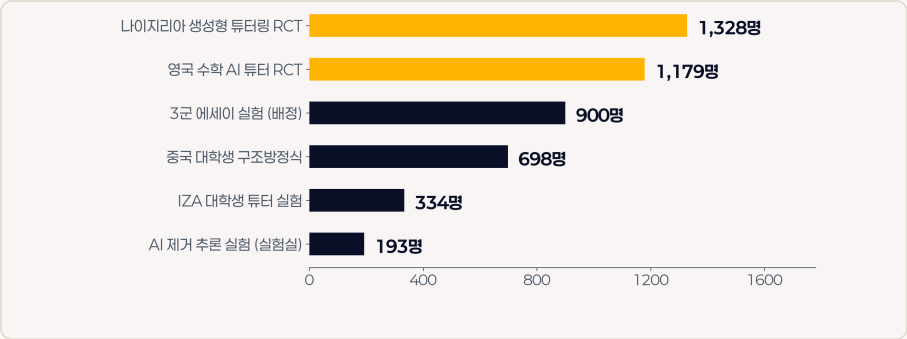
교육 실험 네 건을 같은 축에 놓으면 패턴이 보입니다. 나이지리아 1,328명 실험은 0.31 SD, IZA 대학생 334명은 0.23 SD, 타이베이 10개 고교의 적응형 문제배열은 0.15 SD, Khanmigo 2년 실험은 학기당 1.3 백분위(연 0.06~0.08 SD)였습니다. 그런데 영국 1,179명 실험에서 AI 튜터는 중간 퀴즈에서 표준 수업보다 나았지만(정답률 59.6%→63.5%), 외부 채점 표준화 검사에서는 차이가 사라졌습니다(−0.01~+0.02 SD). 전문가 인간 튜터도 −0.09 SD였습니다.

병목은 접근이 아니라 사용이다

Khanmigo 실험의 가장 중요한 숫자는 효과크기가 아닙니다. 학생의 96%가 한 번 이상 시도했지만, 중앙값 기준으로 연습한 날의 3분의 1에만 메시지를 보냈고, 오답이 나온 연습 세션의 17%에서만 튜터와 대화했습니다. 보낸 메시지도 대부분 단답이나 추천 프롬프트 클릭이었습니다. 접근을 주는 것과 사용하게 만드는 것은 다른 문제입니다. 영국 실험에서는 전문가가 94,385건의 AI 발화 초안 가운데 94.7%를 수정 없이 승인했고, 수학 오류는 1,000건당 3건 미만이었습니다. 품질은 이미 충분하고, 남은 문제는 참여입니다.

VII. 교육 — 증거의 단위가 학교로 옮겨갔다

1,328명·1,179명·900명 — 표본이 커질수록 효과는 설계의 함수임이 분명해진다



자료: 각 논문 공개 초록. 표본은 분석 또는 배정 기준(3군 실험은 900명 배정, 659명 분석)

표본이 커지면 무엇이 보이는가

교실 실험의 표본이 1,000명을 넘어서면서 세 가지가 드러났습니다. 첫째, 효과는 과제 근접성에 반비례합니다. 플랫폼 안의 문제와 중간 퀴즈에서는 효과가 크고, 외부 표준화 검사에서는 사라집니다. 둘째, 조합이 효과를 만듭니다. 나이지리아 연구는 생성형 AI, 교육과정 정렬 활동, 짝 활동, 교사 안내를 12회 90분으로 묶었고, 영어 0.238 SD, AI 지식 0.309 SD, 학교 시험 0.206 SD를 관찰했습니다. 셋째, 디지털 접근 격차가 상쇄 요인입니다. 디지털 접근 상호작용은 −0.182 SD(95% CI −0.370~0.005)로 추정됐습니다.

설계가 결과를 만든다

같은 AI를 쓰고도 효과가 0.02 SD에서 0.31 SD까지 벌어지는 이유는 표본이 아니라 설계입니다. 가장 큰 효과를 낸 조건은 ① 교사가 안내하고 ② 교육과정에 정렬된 과제를 쓰고 ③ 학생끼리 짝으로 작업하며 ④ 12회의 고정 세션을 가진 경우였습니다. 가장 작은 효과를 낸 조건은 ① 도구만 주고 ② 사용 여부를 학생에게 맡긴 경우였습니다. AI 도입 예산의 절반은 도구가 아니라 운영 설계에 배정되어야 한다는 뜻입니다.

VIII. 적용 — 위임을 네 단계로 회계 처리하기

과업 정의 · 맥락 설계 · 검증 설계 · 위임 회계, 그리고 연구가 남긴 네 가지 공백

단계	질문	측정 지표	실패 신호
1. 과업 정의	이 과업은 조회·계산·해석 중 무엇인가	과업 유형별 정답률, 필요 맥락 목록	모든 과업에 같은 프롬프트를 쓴다
2. 맥락 설계	어떤 메타데이터·용어·규칙을 먼저 줄 것인가	사전 매핑 비율, 사후 보충 비율	사후 보충이 30%를 넘는다
3. 검증 설계	누가 어떤 순서로 승인하는가	수정 없이 승인된 비율, 반려 사유 코드	수정률 5% 미만인데 오류율이 줄지 않는다
4. 위임 회계	무엇을 자동화하고 무엇을 남길 것인가	검토 시간, 비보조 성과, 재작업률	AI 제거 후 성과가 기준선 아래로 떨어진다

네 단계를 한 장에 넣는 이유

이번 호의 논문은 각각 다른 층을 다뤘지만 실무에서는 순서가 중요합니다. 과업을 정의하지 않고 맥락을 넣으면 34.9%의 사후 보충이 됩니다. 검증을 설계하지 않고 멀티 에이전트를 도입하면 추적 비용만 늘어납니다. 위임 회계 없이 도구를 보급하면 AI 제거 후 성과가 기준선 아래로 떨어지는 집단을 만듭니다(d=−0.28). 반대로 네 단계를 갖춘 조직은 같은 모델로 더 큰 효과를 냅니다.

조직에 바로 적용하는 세 문장

① 에이전트 도입 전에 과업 목록을 조회·계산·해석으로 분류하고, 유형별로 필요한 맥락을 표로 만듭니다. ② 승인 화면에 "수정 없이 승인" 비율을 표시하고, 그 비율이 90%를 넘으면 검증 설계를 의심합니다. ③ 분기마다 AI를 끈 상태의 성과를 측정해 기준선과 비교합니다. 세 문장은 도구 도입보다 먼저 끝낼 수 있고, 도구 선택의 기준이 됩니다.

연구가 아직 답하지 않은 것

정직하게 남겨야 할 공백이 있습니다. 첫째, 대부분의 교실 실험이 10주~2년입니다. 효과가 누적되는지 사라지는지는 더 긴 관찰이 필요합니다. 둘째, 다수 연구가 단일 국가·단일 과목에서 이뤄졌습니다. 셋째, 인지 위임 연구는 즉시적 근전이(decrement)를 측정했고, 장기적 기능 저하를 입증한 것은 아닙니다. 넷째, 경영 연구의 상당수는 횡단 설문입니다. 인과의 방향은 아직 열려 있습니다. 이 네 가지 공백을 알고 인용하는 것이 실무에서 연구를 쓰는 방법입니다.

적용 맥락을 설계하고, 순서를 정하고, 증거의 단위를 표본으로 확인하라

English Summary / 영문 요약

Context and evidence, not model size, set the boundary of delegation

English Summary

Four fields published in November converge on one finding: outcomes are set by the structure of inputs and the procedure of verification, not by model capability. In the Journal of Intelligence and Information Systems 32(3), Park and Cho show that enterprise LLM analysis agents need different context by task type: simple lookups reach a ceiling with the base schema, metric calculation improves most with extended technical metadata, and business interpretation depends on a semantic layer. Post-hoc context supplementation covered only 34.9% of required context. The same issue compares single and multi-agent financial report generation, and a hierarchical MCP-RAG design for disaster-message decisions; the value of role separation lies in traceability rather than accuracy.

Management, cognition and classrooms

Korea Management Review 55(5) finds that employees AI skills convert into technology commercialisation capability conditional on strategic flexibility, with risk management in the moderated mediation path. A three-arm experiment with 659 analysed participants shows that prior dependent use of AI lowered later unaided reasoning quality by 0.38 on a seven point scale (d = -0.28), while a self-generation-first workflow showed no such deficit. A structural model of 698 students finds perceived AI intelligence raises critical thinking directly and through immersion, but immersion that increases dependency lowers it (beta -0.25). In classrooms, a two-year Khanmigo trial across 18 Tennessee middle schools raised math achievement by 0.06-0.08 standard deviations a year, yet the median student messaged the tutor on only a third of practice days. A 1,179-student UK trial found gains on a midtrial quiz but none on an externally scored test.

Sample size and design

A Nigerian randomised trial of 1,328 students reports 0.310 standard deviations overall, with English at 0.238 and AI knowledge at 0.309. A 334-student IZA experiment reports 0.23, and adaptive problem sequencing in ten Taipei high schools 0.15. The spread is explained by design: teacher guidance, curriculum alignment, peer work and fixed sessions produced the largest effects, while access alone produced the smallest. The practical rule is a four-step ledger: define the task, design the context, design the verification, and audit what remains when the tool is removed.

One line **Context, order and evidence — three levers that cost less than a new model**

발행 CONDERI AI산업전략연구실 | 문서등록번호 CONDERI-AR-2026-11-17 | 문의 contact@conderi.com