CONDERI

ACADEMIC & PAPER RESEARCH | 대외 공개판

대신 해 준 성과, 남지 않은 실력

The Score the AI Earned, the Skill It Left Behind — Seven Studies on Delegation, Verification and Retention

이 리포트가 다루는 것

미국 고등학생 2,456명을 12주간 무작위로 나눈 실험에서, AI 튜터와 함께 푼 학생들의 학습 중 정답률은 72.1%까지 올랐습니다. 그러나 2주 뒤 AI 없이 치른 시험에서는 41.6%로 떨어졌습니다. AI를 한 번도 쓰지 않은 학생은 54.2%였습니다. 같은 주에 게재된 국내외 논문 일곱 편은 같은 질문을 서로 다른 각도에서 검증합니다. 멀티 에이전트는 출처를 늘렸지만 확신은 줄였고, 직장인 303명은 거버넌스가 분명할 때만 AI에 일을 맡겼고, LLM은 행동 규칙을 접지했을 때만 인간의 집단 행동을 흉내 냈습니다. 이 리포트는 기술·경영·과학·교육 네 섹션으로 그 증거를 정리하고, 기업이 쓸 수 있는 위임·검증·잔존의 3층 프레임워크를 제시합니다.

41.6%
AI 튜터 사용군의 2주 후
비보조 시험 점수(AI 없음 54.2%)

61.0%
계층형 멀티 에이전트의
재난문자 문안 승률

40 → 100%
멀티 에이전트 전환 후
출처 명시 충족률

d=1.42
학습 최적화 모델의
비보조 성취 효과크기

I 기술
에이전트를 나누면
무엇이 늘어나는가

II 경영
위임은 신뢰의 함수,
신뢰는 거버넌스의 함수

III 과학
조정은 되지만
행위감은 남지 않는다

IV 교육
세션 점수와
잔존 점수의 역전

Executive Summary / 요약

Seven studies on delegation, verification and retention

41.6%
AI 튜터군 2주 후 비보조 점수

+1.572 SD
AI군 연습 점수(대조군 대비)

34.9%
LLM이 사후 확보한 필요 맥락 비율

231세션
LLM 집단행동 재현 실험 규모

오늘의 결론 — AI는 성과를 만들지만 실력은 남기지 않는다

이번 주 게재된 일곱 편의 논문은 서로 다른 방법으로 같은 결론에 도달했습니다. AI는 과업 수행 점수를 올리지만, 그 점수가 인간의 역량으로 전이되지는 않습니다. 825명·44개 학급을 무작위 배정한 재분석에서 AI 군의 연습 점수는 대조군보다 1.572 표준편차 높았지만, 비보조 시험 점수는 −0.063 표준편차였습니다. 연습 점수가 시험 점수를 예측하는 기울기는 대조군 0.354에서 AI 군 0.093으로 떨어졌습니다. 즉 AI와 함께 받은 점수는 실력을 덜 보여줍니다. 해법도 같은 주에 나왔습니다. 학습 자체를 보상하는 모델은 잔존 점수를 68.9%로 끌어올렸고, 거버넌스와 책임성을 명시한 조직에서만 직원들이 AI에 중요한 일을 맡겼습니다.

① 기술 — 늘어난 것은 설명이다

계층형 멀티 에이전트는 재난문자 문안에서 기존 문안 대비 61.0% 승률을 기록했습니다. 금융투자보고서 통제실험에서는 출처 명시가 40%에서 100%, 위험 고지가 90%에서 100%로 올랐지만, 과도한 확정 표현 억제는 100%에서 80%로 떨어졌습니다.

② 경영 — 위임은 거버넌스가 결정한다

직장인 303명 설문에서 거버넌빌리티와 책임성 명확성이 AI 신뢰의 핵심 선행요인이었습니다. 신뢰는 태도·주관적 규범·지각된 행동통제를 거쳐 위임 의사로 이어졌고, 주관적 규범의 효과가 가장 컸습니다. AI 대리성 지각은 위험을 증폭시켜 신뢰를 약화시켰습니다.

③ 교육 — 점수는 올랐고 실력은 빠졌다

12주 3상 실험에서 GPT-4o 군은 학습 중 72.1%, 비보조 41.6%였습니다. 같은 기간 AI를 쓰지 않은 군은 54.2%였습니다. 선행 12편 메타분석도 단기 +16.8%, 잔존 −19.4%로 같은 방향을 가리켰습니다.

과학 — 인지적 접지가 조건

LLM에게 인간의 행동 규칙을 먼저 주면 집단 행동 재현의 궤적 적합도가 개선됐습니다(231세션, 7개 모델). 반대로 조정만 시키면 공동 과제 수행은 가능하지만 행위감은 인간 파트너보다 약했습니다(N=115).

이 리포트의 사용법

① AI 도입 KPI를 세션 성과에서 지연 성과로 바꾸십시오. ② 에이전트 구조를 바꿀 때 책임성 체크리스트를 함께 측정하십시오. ③ 위임 범위를 정할 때 권한·한계·책임 소재를 문서로 먼저 합의하십시오.

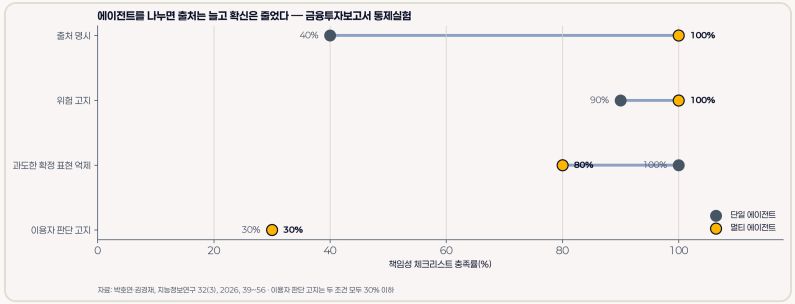
이 리포트의 방법 — 논문을 초록이 아니라 설계로 읽는 방식

게재 정보는 DBpia·KCI·Nature·Frontiers·NBER의 논문 페이지와 초록, 그리고 각 저널의 권·호를 확인해 정리했습니다. 수치는 논문이 제시한 표본·효과크기·충족률을 그대로 인용했고, 본문에서 확인할 수 없는 값은 사용하지 않았습니다. 실험 설계(무작위 배정 여부, 대조군, 반복 횟수)를 함께 적은 이유는 같은 숫자도 설계에 따라 해석이 달라지기 때문입니다.

한 줄 AI는 점수를 만들고, 거버넌스는 위임을 만들고, 접지는 재현을 만든다 — 남는 것은 설계뿐이다

I. 기술 — 멀티 에이전트의 값은 설명에서 나온다

What multi-agent systems actually improve



자료: 박호연·김경재, 지능정보연구 32(3), 2026

에이전트를 나누면 설명은 늘고 확신은 줄어든다

박호연(강원대병원)·김경재(동국대)는 동일한 언어모델과 검색 환경에서 단일 에이전트와 역할 분담형 멀티 에이전트를 통제 비교했습니다. 평가 항목은 처리 시간, 응답 길이, 실행 안정성, 그리고 책임성(RAI) 체크리스트였습니다. 멀티 에이전트는 출처 명시(40%→100%)와 위험 고지(90%→100%)에서 개선됐지만, 처리 시간이 늘고 과도한 확정 표현 억제는 100%에서 80%로 떨어졌습니다. 이용자 판단 고지는 두 조건 모두 30% 이하로, 구조를 바꿔도 해결되지 않는 영역이 남았습니다.

연구	설계	핵심 결과	실무 해석
재난문자 의사결정 지원 (연세대 정보대학원)	MCP·RAG 기반 계층형 멀티 에이전트 + 사람 검증(HITL)	기존 문안 대비 종합평가 승률 61.0%	법규 지식베이스를 붙인 검색 구조가 공공 문서 품질을 끌어올린다
금융투자보고서 생성 (강원대병원·동국대)	3개 모델 × 5회 반복, RAI 체크리스트	출처 40→100%, 위험 고지 90→100%, 확정 표현 억제 100→80%	에이전트 분업은 정보 통합형 책임만 개선한다
멀티 에이전트 SW 교육 (전주교대·한국해양대)	학습자 오케스트레이터 + 전문화 에이전트, 사례 구현	역할 분리가 일관성·완성도를 높임	교육 목표를 도구 사용에서 조정 역량으로 재정의

기술 멀티 에이전트는 답을 고치는 장치가 아니라, 책임의 위치를 드러내는 장치다

II. 경영 — 위임은 거버넌스의 함수다

Why employees delegate to agents, and why context decides output



자료: 이종훈·안현철, 지능정보연구 32(1), 2026

직원이 AI에 일을 맡기는 조건 — 거버넌스와 책임의 명시

이종훈·안현철(국민대)은 신뢰 형성 이론과 계획된 행동이론(TPB)을 통합해 직장인 303명의 응답을 PLS-SEM으로 분석했습니다. 거버너빌리티와 책임성 명확성이 AI 신뢰의 핵심 선행요인이었고, 신뢰는 태도·주관적 규범·지각된 행동통제를 매개로 위임 의사에 영향을 줬습니다. 특히 주관적 규범, 즉 조직 동료들이 AI를 어떻게 보는지가 행동 의도에 가장 큰 영향을 미쳤습니다. AI 대리성 지각은 지각된 위험을 증폭시켜 신뢰를 약화시키는 부정 경로를 만들었습니다.

요인	측정 내용	위임에 미치는 방향
거버너빌리티	AI가 무엇을 할 수 있고 어디서 멈추는지의 명시	신뢰를 높임(핵심 선행)
책임성 명확성	결과에 대한 책임 소재의 명시	신뢰를 높임(핵심 선행)
주관적 규범	동료·조직이 AI 사용을 어떻게 보는가	행동 의도에 가장 큰 영향
AI 대리성 지각	스스로 판단한다고 느끼는 정도	지각된 위험을 키워 신뢰를 낮춤

맥락이 없으면 에이전트는 정답을 못 만든다 — KT 연구

박기준·조미영(KT)은 통신 도메인 합성데이터로 LLM 분석 에이전트의 데이터 맥락 효과를 사전등록 설계로 실험했습니다. 단순 조회는 기본 스키마만으로도 천장 효과를 보였지만, 지표 계산은 확장 기술 메타데이터가, 업무 해석은 시맨틱 레이어가 성과를 갈랐습니다. 실행 규칙과 컬럼 연결을 뺀 자연어 정의 중심의 비즈니스 글로서리는 단독 효과가 제한적이었고, 사후 필요 맥락 포함률은 34.9%에 그쳤습니다. AI-Ready Data는 데이터가 아니라 맥락 인프라라는 결론입니다.

III. 과학 — 조정은 되고, 행위감은 남지 않는다

Cognitive grounding, metacognition and the limits of agency

① 집단 행동의 재현 — 인지적 접지

Zhu·Heydari(노스이스턴대)는 Nature Computational Science(11월 16일)에서 공급망 실험(비어 게임) 231세션으로 7개 LLM의 집단 행동 재현을 검증했습니다. 행동 모델의 질적 규칙을 먼저 주는 "인지적 접지"를 적용하면 궤적 적합도가 모든 모델에서 개선되고 추론된 의사결정 파라미터가 인간 기준 쪽으로 이동했습니다. 다만 개선 폭은 역할과 모델에 따라 달랐습니다.

② 불확실성의 메타인지 — 훈련은 되지만 전이되지 않는다

Steyvers 등(UC 어바인)은 Communications AI & Computing(11월 5일)에서 두 모델을 일반상식·수학·퀴즈로 미세조정했습니다. 확신도와 실제 정답률의 정합성은 훈련 도메인에서 개선됐지만 새 도메인 전이는 제한적이었고, 단일 문항 확신도와 쌍대 비교 과제 사이의 전이는 나타나지 않았습니다. 멀티태스크 훈련이 더 넓은 개선을 만들었습니다.

③ AI와의 협업 기억 — 48시간 뒤에도 남았다

Lei·Liu(Current Psychology, 11월 9일)는 AI로 표시된 파트너와의 협업 회상이 개인 회상보다 더 많은 정답을 남기며, 그 이득이 48시간 뒤에도 유지된다고 보고했습니다. 인간 파트너 조건과 통계적으로 유의한 차이는 없었고, 협업은 주관적 기억 부담의 증가도 막았습니다.

④ 행위감 — 조정은 되지만 책임감은 약하다

Bender 등(11월 9일, Technology, Mind, and Behavior)의 실험(N=115)에서 응답 조정만 요구될 때는 인간과 AI 공동 행위자의 공동표상이 비슷했습니다. 그러나 결과 판단이 더해지면 인간 공동 행위자가 더 강한 공동표상과 대리 행위감을 만들었습니다. AI와 협업할 때 행위감이 가장 약했습니다.

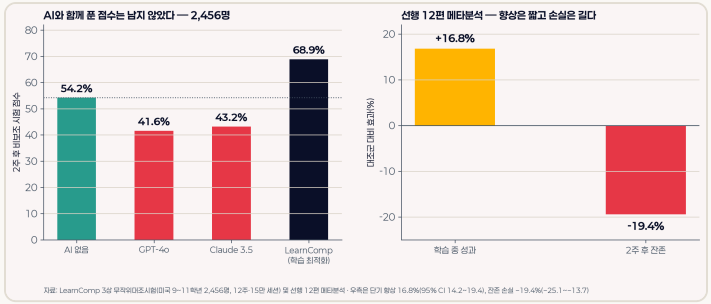
과학 섹션이 기업에 주는 한 문장

LLM은 인간의 판단을 흉내 낼 수 있지만, 그 판단에 대한 책임감을 함께 만들지는 않습니다. 따라서 AI를 "인지 파트너"로 쓰는 설계(기억·탐색·초안)와 "판단 대리자"로 쓰는 설계(승인·처분·평가)를 분리해야 합니다. 앞의 영역은 자동화가 값을 만들고, 뒤의 영역은 사람의 서명이 값을 만듭니다.

과학 조정은 자동화되지만 행위감은 자동화되지 않는다 — 책임의 자리는 비워 둘 수 없다

IV. 교육 — 세션 점수와 잔존 점수는 반대로 움직인다

Assisted performance versus durable learning



자료: LearnComp 3상 무작위대조시험 및 선행 12편 메타분석

무작위 대조시험이 드러낸 역전 — 학습 중 72.1%, 2주 후 41.6%

미국 9~11학년 2,456명을 네 집단(AI 없음, GPT-4o, Claude 3.5, 학습 최적화 모델)으로 나눈 12주 실험에서, 수행 중심 AI군의 세션 정답률은 70~72%까지 올랐지만 2주 뒤 비보조 시험에서는 GPT-4o 41.6%, Claude 3.5 43.2%로 AI를 쓰지 않은 집단(54.2%)보다 낮았습니다. 반면 학습을 보상하도록 훈련한 모델(LearnComp)은 68.9%로 대조군보다 14.7%포인트 높았고, 효과크기는 Cohen's d 1.42였습니다. 같은 논문의 선행 12편 메타분석도 단기 +16.8%, 잔존 −19.4%를 보고했습니다.

진단 가치의 붕괴 — 825명 재분석

Horchani(Frontiers in Education, 10월 22일)는 44개 학급·825명의 2,848개 세션 관측을 재분석했습니다. 연습 점수는 AI군에서 0.931~1.572 표준편차 높았지만 시험 점수는 −0.06~−0.09 표준편차였습니다. 연습→시험 기울기는 대조군 0.354, AI군 0.093~0.123으로 낮아졌습니다. 보조 점수로 실력을 판단하면 안 된다는 뜻입니다.

구조가 값을 만든다 — 6,000명 현장실험

Oreopoulos 등(NBER w35621)은 중학생 6,000여 명을 AI 지원과 전통적 컴퓨터 보조학습으로 나눴습니다. AI군은 더 느리게 진행하고 문항을 덜 풀었지만, 시도에 도달한 경우 정답률이 높았고 오답 이후 다음 시도의 정답률이 개선됐습니다. 숙달 기반 절차와 결합했을 때만 지연 시험의 이득이 나타났습니다.

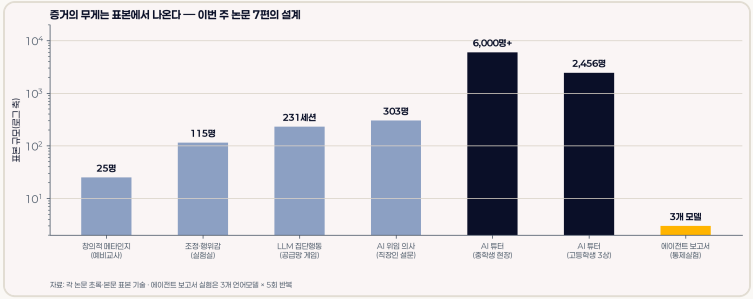
질문 중심 스캐폴딩 — 25명

강다현·이정민(이화여대, 창의력교육연구 26(1))은 예비교사 25명에게 LLM 기반 대화형 에이전트를 적용했습니다. 창의적 메타인지가 유의하게 향상됐고, 특히 문제 이해와 아이디어 생성 단계에서 효과가 컸습니다. 상호작용 로그의 순차 패턴분석에서 점검·조절 활동이 반복 활성화되는 양상이 확인됐습니다.

교육 AI 튜터의 KPI를 세션 정답률에서 2주 후 비보조 성취로 바꾸면, 제품 설계가 통째로 바뀐다

V. 증거 — 표본이 결론의 무게를 정한다

Reading studies by design, not by headline



자료: 각 논문의 표본 기술

연구	표본·설계	핵심 수치	해석의 한계
LeamComp (2026)	2,456명, 4상 무작위, 12주	비보조 68.9% vs 54.2%, d=1.42	단일 교육과정·단일 국가, 12주
Horchani (2026)	825명, 44학급, 재분석	연습 +1.572 SD, 시험 −0.063 SD	2차 분석, 천장효과 민감
NBER w35621 (2026)	6,000명+, 현장 무작위	오답 후 정답률 개선	숙달 절차와 결합 시에만 유의
이종훈·안현철 (2026)	303명, PLS-SEM	거버넌스→신뢰→위임	횡단 설문, 자기보고
Zhu·Heydari (2026)	231세션, 7개 LLM	접지 후 궤적 적합 개선	단일 실험 과제(공급망)
Bender 등 (2026)	N=115, 실험실	결과 판단 시 행위감 격차	실험실 과제, 일반화 제한

증거를 읽는 순서 — 설계, 표본, 그리고 반복

무작위 배정과 대조군이 있는 연구(LearnComp, NBER)는 인과 해석이 가능하지만, 횡단 설문은 관계의 방향을 확정하지 못하고, 재분석 연구는 원자료의 한계를 물려받습니다. 이 리포트가 결론을 하나로 모으지 않고 설계별로 나눠 적은 이유입니다. 인용할 때도 어떤 설계에서 나온 수치인지를 함께 옮기십시오.

VI. 프레임워크 — 위임·검증·잔존의 3층

A three-layer test before scaling any agent

층	질문	설계 요소	측정 지표
1. 위임(Delegation)	무엇을 맡길 것인가	권한 범위, 정지 조건, 승인 단계	위임 후 재작업률, 예외 발생률
2. 검증(Verification)	무엇을 확인할 것인가	출처 명시, 위험 고지, 확정 표현 억제, 사람 검토(HITL)	RAI 체크리스트 충족률, 표본 검토 적중률
3. 잔존(Retention)	무엇이 사람에게 남는가	비보조 과제, 지연 평가, 설명 요구	2주 후 비보조 성취, 오답 교정률

3층 프레임워크 — 도입 심사를 통과하는 질문 순서

첫째 층은 위임입니다. AI가 무엇을 하고 어디서 멈추는지를 계약서 수준으로 적습니다. 둘째 층은 검증입니다. 산출물의 출처·위험·확신 수준을 측정하고, 사람이 어느 비율로 검토할지 정합니다. 셋째 층은 잔존입니다. AI가 대신한 일이 조직의 역량으로 남았는지, 즉 AI 없이도 같은 일을 할 수 있는지를 정기적으로 확인합니다. 세 층을 모두 통과하지 못한 도입은 파일럿에 머물러야 합니다.

교육기관에 적용하면

튜터 도입 심사에서 세션 정답률 대신 2주 후 비보조 평가를 제출하도록 요구하십시오. 교사에게는 AI 사용 로그와 오답 교정 기록을 함께 보게 하십시오. 학습 최적화 모델의 효과크기(d=1.42)는 "학습을 보상하는 설계"에서 나왔습니다.

기업에 적용하면

보고서 자동화는 출처·위험 고지 충족률을 SLA로 넣으십시오. 구조를 멀티 에이전트로 바꿀 때 확정 표현 억제 지표가 함께 떨어지는지 확인하십시오. 위임 전에 권한·한계·책임 소재를 문서로 합의하면 신뢰 형성 비용이 줄어듭니다.

공공에 적용하면

재난·안전 문서는 법규 지식베이스를 검색에 붙인 구조가 품질을 끌어올렸습니다(승률 61.0%). 다만 최종 발송은 사람의 승인을 남겨야 합니다. AI는 초안을 빠르게 만들지만, 발송의 책임은 기관에 남습니다.

프레임워크 위임·검증·잔존을 모두 통과하지 못한 AI 도입은 파일럿에 머물러야 한다

English Summary

For global readers and archive

Overview

Seven studies published this week converge on one finding: AI raises task performance but does not transfer capability to the person. In a four-arm randomised trial of 2,456 US high-school students, students using GPT-4o scored 72.1 percent on assisted practice but 41.6 percent on an unassisted test two weeks later, below the 54.2 percent of students who never used AI. A model trained to reward learning reached 68.9 percent, an effect size of 1.42. A reanalysis of 825 students in 44 classrooms found practice scores 1.572 standard deviations above control while exam scores were 0.063 below, with the practice-to-exam slope falling from 0.354 to 0.093.

Technology and management

A controlled comparison in the Journal of Intelligence and Information Systems found that multi-agent orchestration raised source attribution from 40 to 100 percent and risk disclosure from 90 to 100 percent, but cut overcertainty suppression from 100 to 80 percent; notices that users must judge for themselves stayed below 30 percent in both conditions. A hierarchical MCP-RAG agent for Korean public disaster alerts won 61.0 percent of paired comparisons against existing templates. Among 303 employees, governability and accountability clarity were the key antecedents of trust in agentic AI, and subjective norms had the largest effect on willingness to delegate; perceived agency amplified risk and weakened trust.

Science and implications

A Nature Computational Science study of 231 supply-chain sessions with seven models showed cognitive grounding improves simulation of collective behaviour. An experiment with 115 participants found coordination with an AI partner comparable to a human, but a weaker sense of agency when outcomes had to be judged. Firms should separate AI as a cognitive partner from AI as a decision delegate, and test every deployment on three layers: delegation, verification and retention.

Summary **Assisted performance is not capability — measure what remains without the tool**