CONDERI

ACADEMIC & PAPER RESEARCH INSIGHT | 대외 공개판

강사가 설계한 AI만 성적을 올렸다

Only Instructor-Designed AI Moved the Score — 7,200 Students at 19 Universities

이 리포트가 다루는 것

이번 주 학술 지형의 중심에는 하나의 실험이 있습니다. 19개 대학 7,200명을 3개 학기에 걸쳐 추적한 사전등록 무작위 대조시험에서, 강사가 직접 설계한 AI 도구는 AI 차단 평가에서 0.32 표준편차의 학습 개선을 만들었습니다. 같은 기간 표준화 성적에서는 차이가 나타나지 않았고, 참여도와 만족도도 변하지 않았습니다. 효과는 기술이 아니라 설계에서 나왔다는 뜻입니다. 두 번째 실험은 소크라테스식 되묻기가 개인화된 설명보다 나았다는 결과를, 세 번째는 영국 10개 학교 1,179명의 수학 수업에서 AI 튜터가 근접 학습에서는 전문가를 따라잡았지만 표준화 시험에서는 차이가 없었다는 사실을 보고합니다. 이 번호는 국내 학술지의 최신 논문과 함께 그 경계를 정리합니다.

+0.32 SD
강사 설계 AI
학생 단위 효과

7,200명
19개 대학
3개 학기 추적

64%
소크라테스 튜터
대화로 해결한 오류

94.7%
영국 RCT
사람 검수 무수정 승인

기술
작은 모델의 단계적
사후학습 설계

경영
AI 스킬이 사업화
역량이 되는 조건

과학
응답 지연과
비언어 신호

교육
설계 권한과
평가의 분리

Executive Summary / 요약

Design rights, measurement, and the time lag of learning

7,200명	+0.32 SD	0.00 SD	pp.2011-2045
강사 설계 AI RCT 참여 학생	AI 차단 평가 학습 개선	표준화 시험 개선(영국 RCT)	경영학연구 55(5) AI 스킬 논문

오늘의 결론 — 효과는 설계 권한에서 나오고, 성적은 시간이 필요하다

이번 주 공개된 세 실험은 같은 질문에 서로 다른 층위로 답합니다. 첫째, 누가 설계하는가. 19개 대학에서 강사가 직접 만든 AI 도구는 학생 단위 0.32 표준편차, 학급 단위 0.26 표준편차의 개선을 만들었습니다(95% 신뢰구간 각각 0.06~0.59, 0.02~0.51). 둘째, 어떻게 개입하는가. 성인 400명을 대상으로 한 실험에서 소크라테스식 되묻기는 오류의 64%를 대화로 해결했고, 개인화된 설명보다 0.24 표준편차 높은 성과를 냈습니다. 셋째, 무엇으로 측정하는가. 영국 10개 학교 1,179명의 10주 실험에서 AI 튜터는 근접 학습과 중간 퀴즈에서는 전문가 튜터와 동등했지만, 외부 표준화 시험에서는 표준 보충학습과 차이가 없었습니다(-0.01~+0.02 표준편차).

[기술] 작은 모델의 설계 문제

APJIS 36권 3호는 소형 언어모델을 업무용으로 사후학습시키는 단계적 절차를 설계 관점에서 다룹니다. 추론 능력과 대화 유지의 균형이 설계 변수입니다.

[경영] 스킬이 역량이 되는 조건

경영학연구 55권 5호는 구성원의 AI 스킬이 기술사업화 역량으로 전환되는 조건으로 전략적 유연성과 위험관리를 제시합니다. 스킬만으로는 성과가 되지 않는다는 문제의식입니다.

[과학] 지연과 신호의 심리

AI 에이전트의 응답 지연이 수용 의도에 미치는 영향을 나르시시즘 성향에 따라 해석하는 연구와, 발표 평가에서 비언어 행동을 인식하는 연구가 같은 호에 실렸습니다.

[교육] 평가를 AI 밖으로

세 실험이 공통으로 채택한 것은 AI를 쓰지 않는 평가입니다. 감독·대면 필기시험이 효과 측정의 기준이 됐다는 점이 도입 설계에 직접적인 함의를 줍니다.

이 리포트의 방법 — 사전등록 실험과 국내 학술지를 함께 읽는 이유

해외 실험 세 편은 사전등록(pre-registration)된 무작위 대조시험이고, 효과크기는 표준편차 단위로 보고됐습니다. 국내 논문은 한국경영학회 경영학연구 55권 5호(2026.10)와 한국경영정보학회 Asia Pacific Journal of Information Systems 36권 2·3호(2026.06·09)를 확인했습니다. 실험 연구는 인과를, 학술지 논문은 조건을 제공합니다. 이 리포트는 두 층을 겹쳐 실무 순서를 도출했습니다.

한 줄 효과는 설계에서 나오고, 성적은 시간이 필요하다 — 세 실험이 같은 결론에 도달했다

I. 기술 (Tech) — 모델을 고르는 문제에서 학습 절차를 설계하는 문제로

APJIS 36(3) 2026.09

[기술] 작은 모델을 업무에 맞추는 절차 — APJIS 36권 3호의 설계 연구

Asia Pacific Journal of Information Systems 36권 3호(2026.09)에 실린 박진수·정승호의 "Stepwise Post-Training of Compact Language Models for Business Applications: A Design-Oriented Study of Reasoning and Conversational Retention"(pp.694-724)은 프론티어 모델을 호출하는 방식이 아니라, 작은 모델을 업무에 맞게 단계적으로 사후학습시키는 설계를 다룹니다. 제목이 말하는 두 개의 목표가 핵심입니다. 하나는 추론(reasoning)이고, 다른 하나는 대화 유지(conversational retention)입니다. 이 두 목표는 종종 상충합니다. 추론을 강화하는 학습은 대화의 자연스러움을 떨어뜨리고, 대화 데이터에 맞춘 학습은 논리적 일관성을 잃게 만듭니다.

설계 층위	다루는 문제	현업 적용 지점
데이터	업무 문서·대화 로그의 품질과 분포	사내 로그를 학습 데이터로 쓸 수 있는지의 판단
학습 절차	단계적 사후학습(stepwise post-training)의 순서	한 번에 파인튜닝할지, 단계를 나눌지
평가	추론 정확도와 대화 유지의 동시 측정	업무 정답률과 사용자 재사용률을 함께 기록
비용	토큰 단가가 아닌 완료 업무당 총비용	경량 모델 전환의 손익분기점 계산

같은 호의 또 다른 축 — 생성형 AI 리터러시

같은 호에는 노환호·워런칙·김범수의 "Understanding the Role of Generative AI Literacy in Value-Based Adoption"(pp.534-567)도 실렸습니다. 가치 기반 수용 모델의 틀에서 생성형 AI 리터러시가 인지된 가치와 수용 의도에 어떻게 작용하는지를 다룹니다. 두 논문을 나란히 놓으면 기술 층위의 질문(모델을 어떻게 학습시키는가)과 사용자 층위의 질문(사용자가 무엇을 알아야 쓰는가)이 같은 호에 배치된 이유가 보입니다. 도입률은 모델 성능과 사용자 리터러시의 곱으로 결정됩니다.

같은 호의 다른 논문들 — 플랫폼 거버넌스와 인터페이스 설계

APJIS 36권 3호에는 탕와라곤 등의 "Platform Governance and User Behavior: The Impact of Metacritic's Anti-Review-Bombing Policy"(pp.494-533)와 이주현·오중산의 "Classifying UX-Related Reviews in Vertical Commerce Apps Using the 7C Framework"(pp.725-756)도 실렸습니다. 전자는 플랫폼 정책이 사용자 행동을 어떻게 바꾸는지를, 후자는 사용자 리뷰를 7C 프레임으로 분류하는 방법을 다룹니다. 세 논문을 관통하는 질문은 같습니다. 시스템의 성능이 아니라 시스템이 사용자를 어떻게 규율하는가입니다.

II. 경영 (Management) — 스킬이 역량이 되는 조건은 조직이다

경영학연구 55(5) 2026.10 · pp.2011-2045

[경영] AI 스킬은 그 자체로 성과가 아니다 — 경영학연구 55권 5호

경영학연구 55권 5호(2026.10)에 실린 김현정(칼빈대)·김재명(경희대 국제대학원)의 "구성원의 AI 스킬이 기술사업화역량으로 전환되는 조건: 전략적 유연성의 조절효과와 위험관리의 조절된 매개효과"(pp.2011-2045)는 조직이 흔히 범하는 가정을 문제 삼습니다. AI 교육을 많이 시키면 기술사업화 역량이 올라간다는 가정입니다. 이 연구는 그 전환에 조건이 있다고 봅니다. 조절변수로 전략적 유연성을, 매개 경로로 위험관리를 설정한 구조가 그 조건을 형식화한 것입니다.

독법 1 — 스킬은 자원이고 역량은 결과다

개인의 AI 사용 능력은 자원입니다. 자원이 사업화 역량이 되려면 조직이 자원을 재배치할 수 있어야 합니다. 전략적 유연성이 낮은 조직에서는 교육 투자가 개인 효율로만 남습니다.

독법 2 — 위험관리가 경로다

위험관리는 규제 대응이 아니라 의사결정 장치입니다. AI 산출물의 오류를 누가 어떻게 검증하는지가 정해져 있지 않으면, 현장은 AI를 쓰되 책임은 지지 않는 상태가 됩니다.

독법 3 — 측정은 교육 시간이 아니다

교육 시간과 수료율은 투입 지표입니다. 이 연구가 시사하는 성과 지표는 사업화 과제에 AI가 실제로 투입된 건수와 그 결과의 검증 비율입니다.

같은 시기 해외 문헌의 연결 — 인력 분석 도입의 선행 요인

APJIS 36권 2호(2026.06, pp.292-329)의 "Determinants of Workforce Analytics Adoption: Organizational, Technological, and Environmental Perspectives"는 인력 분석(workforce analytics) 도입 결정을 조직·기술·환경 관점에서 정리합니다. 국내 연구가 개인 스킬에서 조직 역량으로 가는 경로를 다뤘다면, 이 논문은 조직이 분석 체계를 도입하는 조건을 다룹니다. 두 연구를 겹치면 실무 순서가 나옵니다. 인력 데이터의 정의와 품질이 먼저이고, 그 위에 스킬·역량 측정이, 마지막에 자동화가 놓입니다.

데이터의 소유권이 먼저다 — 같은 호의 사례 연구

경영학연구 55권 5호에는 이대성의 "조직 간 자원 보완성을 통한 데이터 2차 사용의 이중 가치 창출과 자산화: 유통·물류 산업의 데이터 마찰과 정착 경로 분기에 관한 단일 사례 연구"(pp.1817-1853)도 실렸습니다. 같은 호에 AI 스킬 논문과 데이터 2차 사용 논문이 나란히 있다는 것은 우연이 아닙니다. AI 역량의 전제는 데이터이고, 데이터의 전제는 조직 간 권한입니다. 교육 투자보다 데이터 계약이 먼저라는 순서가 여기서 확인됩니다.

경영 AI 교육의 성과는 수료율이 아니라 사업화 과제에 투입된 건수로 측정해야 한다

III. 과학 (Science) — 지연은 신호가 되고, 비언어는 데이터가 된다

경영학연구 55(5) · J Sci Educ Technol(2026-11-12)

[과학] 응답 지연과 비언어 신호 — AI 평가의 두 경계

경영학연구 55권 5호에는 인간-AI 상호작용의 미시적 조건을 다룬 두 편이 함께 실렸습니다. 정혁진·김한구의 "AI 에이전트의 제품 추천 과정에서 응답 지연이 수용 의도에 미치는 영향: 나르시시즘 성향에 따른 시간적 단서의 해석을 중심으로"(pp.1855-1882)는 같은 답이라도 언제 도착하는지가 수용을 바꾼다는 문제를 다룹니다. 사람은 응답 시간을 능력의 신호로 읽고, 그 해석은 자기 성향에 따라 달라집니다. 김준석·윤여훈·이새롬의 "Speaker-Independent Recognition of Negative Nonverbal Behaviors for AI-Based Oral Presentation Assessment"(pp.1901-1937)는 발표 평가에서 화자와 무관하게 부정적 비언어 행동을 인식하는 문제를 다룹니다.

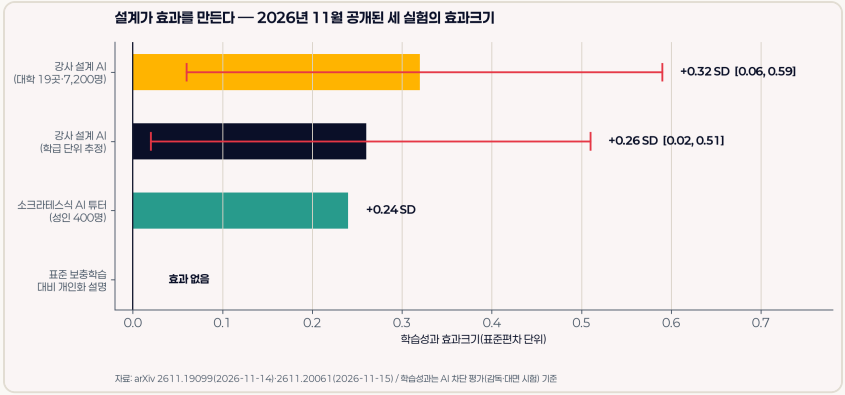
연구	문제	실무 적용
응답 지연과 수용 의도	지연이 능력의 신호로 해석되는 경로	AI 서비스의 응답 시간을 기능이 아니라 UX 설계 변수로 관리
발표 평가의 비언어 인식	화자별 차이를 제거한 부정 신호 탐지	발표·면접 평가 자동화의 공정성 기준 설정
과학 교육의 GenAI 상호작용	텍스트 지원과 심층 참여가 성과에 미치는 관계	도움의 양보다 사고 과정의 흔적을 남기는 설계

세 번째 축 — 과학 수업에서 무엇이 성과를 만드는가

Journal of Science Education and Technology가 11월 12일 공개한 디트리히 등의 연구는 생태학 맥락에서 생성형 AI로 학습할 때 텍스트 지원과 심층 참여가 성과와 어떻게 연결되는지를 다룹니다. 핵심은 도움의 양이 아니라 상호작용의 깊이입니다. 이는 11월 19일 공개된 국내 조사와 정확히 같은 지점을 가리킵니다. AI를 쓰는 학생이 76.8%인데 그대로 옮겨 쓰는 비율이 70.5%라는 사실은, 도구의 보급이 아니라 상호작용의 질이 설계 대상이라는 뜻입니다.

IV. 교육 (EduTech) — 세 실험, 하나의 경계

Instructor-designed AI, Socratic tutoring, and standardised tests



자료: arXiv 2611.19099·2611.20061(2026-11)

연구	설계	핵심 결과
강사 설계 AI (arXiv 2611.19099, 2026-11-14)	사전등록 RCT, 19개 대학 7,200명, 3개 학기	학생 단위 +0.32 SD [0.06, 0.59] 학급 단위 +0.26 SD [0.02, 0.51] 참여도·만족도 변화 없음
소크라테스식 AI 튜터 (arXiv 2611.20061, 2026-11-15)	사전등록 온라인 실험 성인 400명	오류 64%를 대화로 해결 +0.24 SD(설명 조건 대비) 시간은 약 2배, 불만은 더 낮음
영국 수학 AI 튜터 (Google·Eedi, 2026-10-29)	사전등록 RCT, 10개 학교 1,179명, 10주	2차 진단 정답률 43.4%→56.9% 중간 퀴즈 59.6%→63.5% 표준화 시험 −0.01~+0.02 SD

숫자를 읽는 세 가지 주의

첫째, 0.32 표준편차는 큰 효과입니다. 표준편차 1을 한 학기 학습량의 분포 폭이라고 보면, 0.32는 상위 12%가량을 한 단계 끌어올리는 크기입니다. 둘째, 신뢰구간을 함께 봐야 합니다. 학급 단위 추정의 하한이 0.02로 0에 가깝다는 것은 학급 구성에 따라 효과가 사라질 수 있다는 뜻입니다. 셋째, 표준화 시험의 0은 실패가 아니라 시차입니다. 10주 동안 개념을 깊게 다루면 범위를 넓게 훑는 표준 보충학습보다 진도가 느려질 수 있습니다.

V. 교육 (EduTech) — 설계 권한이 효과의 크기를 정한다

What the three experiments actually established

[교육] 무엇이 효과를 만들고, 무엇이 아직 효과를 만들지 못했는가

세 실험을 겹쳐 읽으면 설계 원칙 네 가지가 남습니다. 첫째, 설계 권한을 현장에 두는 것이 효과의 조건입니다. 강사가 자기 수업과 학생을 알기 때문에 만든 도구가 0.32 표준편차를 만들었고, 중앙에서 일괄 배포한 도구는 효과가 없었습니다. 둘째, 되묻는 개입이 설명하는 개입보다 낫습니다. 소크라테스식 튜터는 개인화된 설명보다 0.24 표준편차 높았고, 개인화 설명은 일반 설명과 차이가 없었습니다. 셋째, 표준화 성적은 아직 움직이지 않았습니다. 10주 실험에서 AI 튜터와 표준 보충학습의 차이는 사실상 0이었습니다. 넷째, 만족도와 학습은 다른 지표입니다. 강사 설계 AI는 참여도와 만족도를 바꾸지 않았지만 학습은 개선했습니다.

원칙 1 — 설계 권한

도구를 고르는 권한과 설계하는 권한은 다릅니다. 교사가 문항·피드백·순서를 조정할 수 있어야 효과가 생깁니다. 반대로 중앙 배포형 도구는 효과가 관측되지 않았습니다.

원칙 2 — 되묻기 우선

정답과 설명을 주는 기능보다, 학생이 스스로 고치도록 유도하는 대화가 효과적입니다. 다만 시간이 두 배 걸린다는 비용을 함께 계산해야 합니다.

원칙 3 — 평가의 분리

효과 측정은 AI를 쓰지 않는 평가로 해야 합니다. 감독·대면 필기시험과 외부 표준화 시험이 기준이 됐을 때, 근접 학습과 전이 학습의 차이가 드러납니다.

한국 교육 현장에 옮기면 — 기준을 먼저 만들고 도구를 나중에 붙인다

경기도교육청이 11월 9일 공청회에서 제시한 학교급별 기준안은 이 순서와 일치합니다. 초등은 교사 감독 아래 검증된 교육용 AI만, 중등은 교사가 정한 범위 안에서, 고등은 출처와 이용 내역 공개를 전제로 허용하는 3단 구조입니다. 도구를 먼저 보급하고 기준을 나중에 만드는 순서를 뒤집은 것입니다. 다만 세 실험이 공통으로 보여준 것은 기준만으로는 부족하다는 점입니다. 기준은 상한을 정하고, 설계 권한은 하한을 만듭니다.

교육 중앙 배포형 도구는 효과가 없었고, 강사가 설계한 도구는 0.32 표준편차를 만들었다

VI. 적용 — 16주 실행 프레임워크

From evidence to a documentable design

단계	할 일	산출물
1~2주	AI 차단 평가 설계(감독·대면 필기)	평가 문항과 채점 기준
3~4주	교사 설계 권한 범위 정의(문항·피드백·순서)	설계 권한표
5~8주	되묻기 중심 튜터 프로토타입 2개 반 운영	대화 로그와 오류 해결률
9~12주	근접 학습·전이 학습 분리 측정	내부 평가와 외부 시험 비교표
13~16주	도입 판단(확대·수정·중단)	의사결정 메모

측정 지표를 세 층으로 나눈다

첫째 층은 근접 학습입니다. 방금 배운 문제를 다시 푸는 능력이고, AI 튜터가 가장 큰 효과를 보이는 지점입니다. 둘째 층은 전이 학습입니다. 4주 뒤 중간 퀴즈처럼 범위가 넓어진 평가이고, 영국 실험에서는 개인화 맥락을 준 AI 튜터가 63.5%로 표준 보충학습(59.6%)을 앞섰습니다. 셋째 층은 표준화 성적입니다. 10주 실험에서는 차이가 나타나지 않았습니다. 세 층을 구분하지 않으면 근접 학습의 성과를 전체 성과로 보고하게 됩니다.

교사에게 필요한 것은 도구가 아니라 설계 시간

세 실험의 공통 변수는 교사·강사의 시간이었습니다. AI가 초안을 만들 수는 있지만, 어떤 오류를 남기고 어떤 질문을 던질지는 수업을 아는 사람이 정해야 합니다. 예산 배분에서 도구 라이선스와 설계 시간을 같은 항목으로 보는 관행이 효과의 크기를 결정합니다.

적용 도구는 값싸고 설계는 비싸다 — 효과의 크기는 비싼 쪽에서 결정된다

English Summary / 영문 요약

Three 2026 experiments on AI in learning

Instructor-designed AI improved learning in a 19-university trial

A pre-registered randomised controlled trial across 7,200 students at 19 universities over three semesters tested whether AI designed by frontline instructors improves learning. Learning was measured with uncurved scores from AI-proof assessments, typically proctored, in-person, pen-and-paper examinations, using a pre-registered difference-in-differences design. The student-level treatment effect was 0.32 standard deviations (standard error 0.13, p = 0.019, 95% confidence interval 0.06 to 0.59) and the class-level estimate was 0.26 standard deviations (standard error 0.12, p = 0.038, confidence interval 0.02 to 0.51). The authors found no change in student engagement, measured by course activity records, or satisfaction, measured by course evaluations.

Socratic tutoring beats explanation

In a pre-registered online experiment with 400 adults practising constraint equations, participants who received Socratic tutoring resolved 64% of their errors through dialogue and scored 0.24 standard deviations higher on practice-aligned posttest questions than those who received generic or personalised explanations. Socratic participants found the session more interesting, less frustrating and less costly, although it took nearly twice as long. The study found no evidence that personalised explanations outperformed generic ones, and neither reliably outperformed simply seeing the correct answer.

A UK mathematics trial and Korean journals

A pre-registered trial with the Eedi platform and Gemini across ten UK state secondary schools followed 1,179 students over ten weeks. Students receiving AI tutoring corrected a mistake 56.9% and 56.5% of the time depending on context, against 43.4% for standard remediation and 57.7% for expert human tutoring. On the mid-trial quiz they scored 63.5% with personalised AI tutoring against 59.6% for standard remediation. On the externally scored Star Maths assessment, AI tutoring showed effects of -0.01 and +0.02 standard deviations relative to standard remediation. Supervising tutors approved 94.7% of 94,385 drafted messages without changes, and audits found mathematical errors in fewer than three in every 1,000 AI messages. Korean journals in the same period examined how AI skills convert into technology commercialisation capability (Korea Management Review 55(5), pp.2011-2045), how response latency affects acceptance of AI recommendations (pp.1855-1882), and speaker-independent recognition of negative nonverbal behaviour for presentation assessment (pp.1901-1937).

Summary **Design authority determines effect size, and standardised scores lag behind practice gains**

발행 CONDERI AI산업전략연구실 | 문서등록번호 CONDERI-AR-2026-11-20 | 문의 contact@conderi.com