CONDERI

ACADEMIC & RESEARCH INSIGHT | 대외 공개판

AI는 누가 설계할 때 효과가 있는가

Who Designs the AI? Evidence from 19 Universities, 7,200 Students and 89 Studies

이 리포트가 다루는 것

2026년 11월 셋째 주에 공개된 네 편의 연구가 같은 질문에 서로 다른 각도로 답했습니다. 교수가 직접 설계한 AI는 19개 대학 7,200명의 성적을 0.32표준편차 올렸지만 만족도는 바꾸지 못했습니다. 소크라테스식 되묻기는 학생이 스스로 오류의 64%를 고치게 했고, 개인화된 설명은 일반 설명보다 나았습니다. 89편을 모은 체계적 검토는 효과가 13주 이상 이어질 때 더 컸다고 보고했습니다. 기술·경영·과학·교육 네 섹션으로 나눠 읽습니다.

+0.32 SD
교수 설계 AI
학습 성과

7,200명
19개 대학
사전등록 RCT

64%
되묻기로
스스로 고친 오류

89편
지능형 튜터링
체계적 검토

I 교육
교수가 만든
AI가 이겼다

II 기술
설명보다
다시 풀기

III 과학
효과는
기간에 비례

IV 경영
AI 스킬은
언제 자산이 되나

i. 요약 — 설계자가 바뀌면 결과가 바뀐다

Design authority, not model size, moved the numbers

0.32 SD	+0.24 SD	83/89	89%	6%
학습 성과 학생 수준 효과	소크라테스 튜터링 효과	긍정 방향 연구 수	13주 이상 긍정 효과 비율	EBIT 5% 이상 AI 기여 기업

한 문장

AI의 효과는 **모델의 성능**보다 **설계자의 위치**와 **개입의 방식**에서 갈렸습니다. 교실에서는 교사가, 실무에서는 인계 지점을 정한 팀이 이겼습니다.

네 편이 공통으로 말한 것

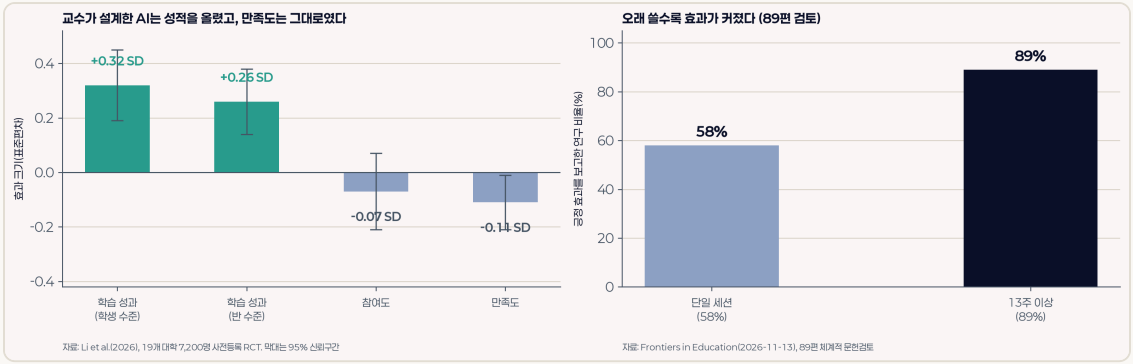
설계 권한을 현장에 두고(교육), 설명 대신 다시 풀게 하고(기술), 충분한 기간을 두고(과학), 역할과 검증 지점을 다시 짜야(경영) 성과가 납니다. 넷 다 **도구 도입**이 아니라 **과정 재설계**의 문제입니다.

이 문서를 읽는 방법

이 리포트는 2026년 11월 13~18일 공개된 논문과 국내 등재지 최신호를 교차 확인해 작성했습니다. 모든 수치는 원문에 표시된 값이며, 실험 설계의 한계도 함께 적었습니다. 해석과 판단은 콘데리 AI산업전략연구실의 것이며 투자 권유가 아닙니다.

I. 교육 — 교수가 만든 AI가 이겼다

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



자료: Li et al.(2026), 19개 대학 7,200명 사전등록 무작위 대조실험. 막대는 95% 신뢰구간

실험 설계

2025년 봄부터 2026년 봄까지 세 학기 동안, 19개 대학의 응용 과학·기술·공학·수학(STEM) 수업 55개(교수-강좌 쌍)에 등록한 7,200명을 대상으로 사전등록 무작위 평가가 진행됐습니다. 처치군 교수들은 자신의 수업에 맞는 AI를 직접 설계했습니다. 19개 반은 소크라테스식 질문을 쓰는 개별 코칭 에이전트를, 18개 반은 목표 연습과 시험 대비를 돕는 튜터를, 25개 반은 학습 관리 시스템을 보완하는 과정 안내 도구를 만들었습니다.

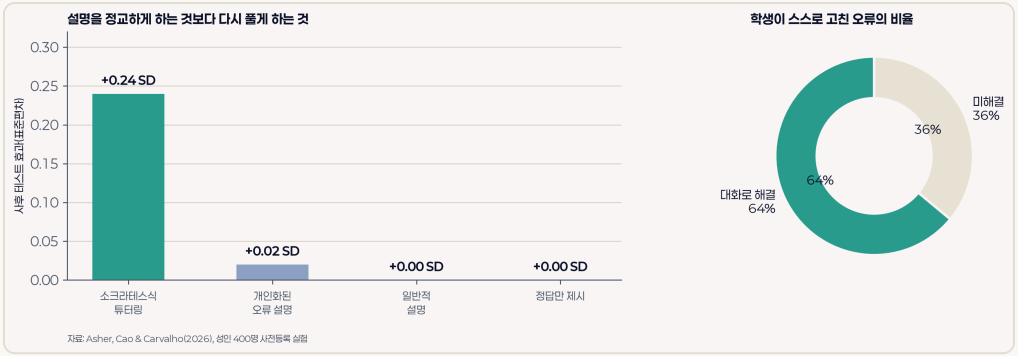
결과 — 성적은 올랐고, 만족도는 그대로였다

학습 성과는 감독·대면 필기시험처럼 **AI가 도울 수 없는 평가**로 측정했습니다. 학생 수준 효과는 0.32표준편차(표준오차 0.13, P=0.019, 95% 신뢰구간 0.06~0.59), 반 수준 효과는 0.26표준편차(P=0.038)였습니다. 반면 참여도는 −0.07표준편차(P=0.60), 만족도는 −0.11표준편차(P=0.29)로 통계적으로 유의한 변화가 없었습니다. 학습량이나 체감이 아니라 **실제 성취**가 움직였다는 뜻입니다.

교육 학생이 좋아했다는 증거 없이 성적이 올랐다. 교육에서 만족도는 성과의 대리지표가 아니다

II. 기술 — 설명을 늘리는 대신 다시 풀게 하라

Scaffolded error correction beats personalised explanation



자료: Asher, Cao & Carvalho(2026), 성인 400명 사전등록 온라인 실험

설계 비교 — 네 가지 처치

참가자 400명은 제약방정식 세우기를 연습하면서 매 시도 뒤 네 가지 중 하나를 받았습니다. 정답만 보기, 일반적 설명, 개인화된 LLM 오류 설명, 소크라테스식 튜터링입니다. 소크라테스식 튜터링은 학생이 스스로 답에 도달하도록 되묻고, 막히면 단계를 낮춰 다시 질문합니다.

결과 — 되묻기가 이겼다

소크라테스식 튜터링을 받은 집단은 대화를 통해 **오류의 64%**를 스스로 해결했고, 연습과 정렬된 사후 문항에서 다른 집단보다 0.24표준편차 높은 점수를 냈습니다. 세션은 약 두 배 길었지만 참가자들은 더 흥미롭고 덜 답답하며 인지적 비용이 낮다고 평가했습니다. 반면 개인화된 오류 설명은 일반 설명보다 나은 증거가 없었고, 두 설명 모두 정답만 본 조건을 안정적으로 앞서지 못했습니다.

설계 함의

정교한 설명을 생성하는 능력은 이미 충분합니다. 부족한 것은 **학생이 다시 시도하도록 붙잡아 두는 구조**입니다. 제품 관점에서 보면 튜터의 핵심 기능은 정답 생성이 아니라 오류 감지, 단계 조절, 종료 조건입니다. 세 가지를 설계하지 않으면 모델을 바꿔도 결과는 달라지지 않습니다.

III. 과학 — 효과는 기간과 수준에 따라 부호가 바뀐다

Duration and expertise decide whether AI helps or hurts

연구	설계	핵심 발견
지능형 튜터링 체계적 검토 (Frontiers in Education, 2026-11-13)	2015-2026년 Scopus·Web of Science·ERIC에서 89편 선별. 사용 중 성과 데이터로 지시를 조정하는 시스템만 포함하고, 시스템을 쓰지 않은 대조군과 비교한 연구만 채택	83편(0.93)이 시스템에 유리한 방향, 71편(80%)이 유의한 우위. 대조군이 우세한 연구는 없음. 단일 세션 연구의 58%가 긍정 효과를 보고한 반면 13주 이상 연구는 89%
전문성 역전과 생성형 AI (Smart Learning Environments, 2026-11-18)	2026년 미국 연구중심 대학 3학년 컴퓨터과학 학생을 대상으로 8주 실험 4회. 지원 수준, 과제 복잡도, 접근 조건을 조작	상세한 안내는 초보 학습자의 외재적 인지부하를 낮추고 성과를 올렸지만, 숙련 학습자에게는 부하를 높이고 성과를 떨어뜨리는 역전이 나타남. 개인화 접근은 초보에게 더 유리
AI 스킬과 기술사업화 (경영학연구 55(5), 2026-10)	구성원의 AI 스킬이 기술사업화역량으로 전환되는 조건을 전략적 유연성의 조절효과와 위험관리의 조절된 매개효과로 검증	AI 스킬 자체는 성과가 아니라 **전환 조건**이 필요. 전략적 유연성이 높을 때 전환 효과가 커지고, 위험관리가 그 경로를 매개합니다

과학 섹션의 결론

세 연구를 겹치면 하나의 규칙이 나옵니다. AI 지원의 효과는 **기간**과 **학습자 수준**에 따라 부호가 바뀝니다. 짧게 쓰면 효과가 작고, 오래 쓰면 커집니다. 초보에게는 상세한 안내가, 숙련자에게는 덜 간섭하는 설계가 맞습니다. 그래서 조직은 하나의 AI 도구를 전사에 배포하기 전에 두 가지를 정해야 합니다. 누구에게 어떤 강도의 안내를 줄지, 그리고 측정 기간을 얼마로 잡을지입니다.

과학 효과는 도구에 붙어 있지 않다. 기간과 수준에 따라 부호가 바뀐다

IV. 경영 — 근거는 늘고 확신은 줄어든다

More sources, less certainty: what the papers measured

논문·자료	표본·방법	실무 시사점
AI 에이전트 응답 지연과 수용 의도 (경영학연구 55(5), 정혁진·김한구)	AI 에이전트의 제품 추천 과정에서 응답 지연이 수용 의도에 미치는 영향을 나르시시즘 성향에 따른 시간적 단서 해석으로 분석	같은 지연도 사용자 성향에 따라 다르게 해석됩니다. 지연을 숨기기보다 **무엇을 하는 중인지 보여주는 설계**가 수용도를 지킵니다
단일·다중 에이전트 금융보고서 생성 비교 (지능정보연구 32(3), 박호연·김경재)	모델·검색 환경을 고정하고 역할 분담형 멀티에이전트와 단일 에이전트를 처리시간, 응답 길이, 실행 안정성, 책임 있는 AI 체크리스트로 비교	출처 명시율 40%→100%, 위험 고지율 90%→100%로 올랐지만 과도한 확정 표현 억제율은 100%→80%로 하락. 이용자 판단 고지 충족률은 두 구조 모두 30% 이하
AI 원가의 재구성 (McKinsey, 2026-11-23)	자사와 소비재·에너지 기업 4곳 사례. 토큰 사용의 멱법칙(상위 10% 사용자가 전체 토큰의 65%), 파일럿→양산 시 지출 4배	모델 라우팅·캐싱·토큰 최적화로 50~60% 절감. 에너지 기업은 통제 계층 도입으로 AI 지출의 약 18% 최적화 기회를 확인

경영 섹션의 세 문장

첫째, 에이전트를 늘리면 근거는 늘고 확신은 줄어듭니다. 출처 표기는 좋아지지만 단정은 거세지지 않습니다. 둘째, 실패는 단계 안이 아니라 **인계 지점**에서 생깁니다. 한국경영정보학회 2026 추계학술대회에서 발표된 120여 편 가운데 여러 편이 같은 결론을 냈습니다. 셋째, 원가는 모델 단가가 아니라 **과제당 비용**으로 관리해야 합니다. 같은 모델도 단계와 토큰 사용에 따라 청구서가 달라집니다.

한국 기업에 옮기면

AI 도입률 89%와 영업이익 기여 37% 사이의 52%포인트를 메우는 것은 모델 교체가 아니라 프로세스 재설계입니다. 이번 주 논문들이 가리키는 실행 순서는 같습니다. ① 인계 시 넘길 필드를 목록으로 정하고, ② 검증을 자동 평가와 사람 확인의 2단으로 두고, ③ 책임을 직책으로 지정하고 로그에 남깁니다.

V. 실무 적용 프레임워크

Five decisions and three numbers

결정	질문	측정 지표
① 설계 권한	현장 전문가가 직접 설계하고 조정할 수 있는가	현장 수정 비율, 수정 후 성과 변화
② 개입 방식	설명을 주는가, 다시 시도하게 하는가	오류 자체 해결률, 세션 완주율
③ 기간 설계	몇 주 동안 반복 사용되는가	4주·13주 시점 효과, 사용 중단율
④ 수준 적합성	초보와 숙련자에게 같은 안내를 주고 있는가	수준별 성과 격차, 인지부하 자기보고
⑤ 원가 구조	토큰당 비용인가 과제당 비용인가	과제당 원가, 자동 라우팅 비율

평가를 설계하는 순서

도입 심사에서는 정확도보다 세 숫자를 먼저 봅니다. ① 출처가 붙은 문장의 비율, ② 인계 지점에서 걸러진 오류 건수, ③ 사람이 최종 확인한 비율입니다. 세 숫자가 오르지 않으면 모델을 바꿔도 결과는 같습니다. 여기에 네 번째로 **과제당 원가**를 붙이면, 성과와 비용을 같은 표에서 비교할 수 있습니다.

프레임워크 좋은 AI 도입은 좋은 모델 선택이 아니라 좋은 실험 설계다

English Summary

For global readers and archive

English Summary

Four studies published between November 13 and 18, 2026 converge on one finding: the effect of AI depends less on the model than on who designs it and how it intervenes. A pre-registered randomised trial across 55 STEM classes at 19 universities and 7,200 students found that AI designed by frontline instructors improved learning by 0.32 standard deviations at the student level (P=0.019) and 0.26 at the course-section level, measured with proctored, AI-proof examinations. Engagement and satisfaction showed no significant change.

Scaffolding beats explanation

In a pre-registered experiment with 400 adults, Socratic tutoring resolved 64 percent of errors through dialogue and raised post-test scores by 0.24 standard deviations relative to generic or personalised explanations, although sessions took nearly twice as long. Personalised LLM explanations showed no reliable advantage over generic ones.

Duration and expertise

A systematic review of 89 intelligent tutoring studies found 83 favouring the system, 71 of them significantly, with positive effects in 58 percent of single-session studies versus 89 percent of studies lasting 13 weeks or more. A separate set of four eight-week experiments found an expertise reversal: detailed guidance helped novices but increased extraneous load and reduced performance for advanced learners. In management research, multi-agent structures raised source attribution from 40 to 100 percent while suppressing overconfident phrasing fell from 100 to 80 percent, and McKinsey reports that AI spending quadruples from pilot to production, making cost per task the governing metric.

Takeaway **Design authority, intervention design and measurement horizon decide the outcome — in classrooms and in firms**