CONDERI

CONDERI ACADEMIC INSIGHT | 학술·연구 인사이트 | 대외 공개판

수행의 증명: 에이전트는 과업을 넘어 프로젝트로, 학습자는 접근을 넘어 참여로

Proof of Performance: Agents Move from Tasks to Projects, Learners from Access to Participation

이 리포트를 읽는 방법

2026년 11월 7일부터 10일 사이 공개된 국내외 논문 아홉 편과 산업 실증 네 건을 기술·경영·과학·교육 네 섹션으로 나눠 문제의식, 방법론, 수치, 실무 시사점 순으로 정리했습니다. 초록 요약이 아니라 **"그래서 무엇을 해야 하는가"**에 답하는 데 초점을 두었습니다. 수치는 논문이 보고한 값 그대로이며, 해석은 콘데리 연구실의 판단임을 밝힙니다.

문서 정보 / Document Information

문서등록번호 Document No.	CONDERI-AR-2026-11-10
발행일자 Published	2026. 11. 10. (화) Tuesday
권호 Volume	VOL. 2026-11-10
분석기간 Coverage	2026. 11. 07 ~ 11. 10 (최근 공개 논문·실증)
수록 논문 Papers	9편 + 산업 실증 4건 (기술 3 · 경영 3 · 과학 2 · 교육 4)
배포구분 Distribution	Public | 일반공개
발행기관 Publisher	CONDERI AI산업전략연구실

65.0 / 100
DeliverBench 최고 성능 (7개 모델·하네스)
90점 기준 통과 과업 50.0%

18.5%
Multi-step SWE 최고 해결률
69개 과업 중 38개는 전 모델 미해결

-0.37 SD
교과 통합 AI 튜터의 최종 성적
2,379명 대학 RCT

0.037
AI 통합→운영성과의 총 간접효과
231명 관리자, '작지만 유의'

한 주의 결론: 능력은 측정되고, 성과는 이어진다

에이전트 논문 여섯 편과 조직·교육 실증 일곱 건이 같은 곳을 가리킨다 — '한 번의 정답'이 아니라 '이어지는 수행'

65.0
DeliverBench 최고 점수. 과업이 아니라 프로젝트를 채점하면 성능이 급락한다

18.5%
Multi-step SWE 최고 해결률. 앞 단계의 오류가 다음 단계로 상속된다

2.97 $/$
ServerBench 최고 기록. 같은 GPU 비용으로 4.87배 토큰 가치

+1.3
Khanmigo 2년 RCT의 학기당 백분위 상승. 참여가 낮으면 효과도 낮다

1. 기술: '과업의 합'은 '프로젝트'가 되지 않는다

이번 주 arXiv에는 에이전트 평가 논문이 몰렸습니다. SlackBench는 실제 연구 프로젝트를 재현해 140개 질의를 냈고, 33개 모델·하네스 가운데 최고가 약 80%, 대부분이 66% 미만이었습니다. DeliverBench는 11개 프로젝트·150개 하위과업을 2~15회의 자료·피드백 라운드로 연결했고, 최고 시스템도 100점 만점에 65.0점이었습니다. Multi-step SWE는 69개 과업에서 최고 해결률 18.5%를 기록했습니다. 공통 결론은 하나입니다 — **능력은 과업 단위에서 측정되지만, 가치는 프로젝트 단위에서 만들어진다.**

해법의 단서도 같은 주에 나왔습니다. Flex-RSI는 에이전트가 '답'이 아니라 **하네스(프롬프트·도구·메모리·제어 루프)**를 스스로 고칠 때 성능이 더 오른다고 보고합니다. ServerBench는 서빙 시스템을 에이전트가 설계하도록 했을 때 GPU 비용 1달러당 지연 보정 토큰 가치가 vLLM 기준의 4.87배에 이르는 사례를 확인했습니다.

2. 조직: 도입이 아니라 배치가 성과를 만든다

국내 실증 세 편은 같은 메시지를 다른 각도에서 확인합니다. 생성형 AI는 **활용 유형**(텍스트 vs 멀티모달)에 따라 성과 경로가 다르고, AI 통합은 **협업(β=0.406)과 프로세스 유연성(β=0.375)**을 거쳐야 민첩성이 되며, 770개 스페인 기업에서는 **배치의 깊이와 넓이의 상호작용**이 성과를 설명합니다. AI 숙련도는 그 자체로 성과가 아니라 **활용 비중과 자율성 변화**를 순차 매개로 성과가 됩니다.

3. 과학: 측정의 정밀도가 예산을 바꾼다

캐시된 실행 흔적만으로 시스템 성능을 예측하는 방법이 문항반응이론(IRT)과 결합해 평균절대오차를 0.13→0.08(SWE-bench Verified), 0.14→0.12(Terminal-Bench 2.0)로 낮췄습니다. 단 한 개의 무작위 과업으로도 유효합니다.
WeGenEditBench는 2,740개 인스턴스·148개 과업 유형에서 항목별 체크리스트로 감점을 먼저 계산합니다. 평가 비용이 곧 의사결정 비용인 시대입니다.

4. 교육: 접근성은 이미 무료, 남은 것은 참여

18개 중학교 2년 RCT에서 Khanmigo는 학기당 1.3백분위를 올렸지만, 중앙값 학생은 연습일의 1/3에만 메시지를 보냈습니다. 2,379명 대학 RCT에서는 코스에 통합된 AI 튜터가 성적을 0.37SD 낮췄습니다. 반면 도쿄대 300명 반나절 RCT에서 '질문을 유도하는' 가이드형 학습은 5.3%p 대비 8.1%p의 이득을 냈습니다. 설계가 결과를 가릅니다.

논문 지도 · 기술 6편
벤치마크가 과업에서 프로젝트로 이동 — Flex-RSI·SlackBench·ServerBench·DeliverBench·Multi-step SWE·TB-MathModel

경영 3편 · 과학 2편
배치의 깊이×넓이, 팀 학습, 숙련의 순차 매개. 측정은 IRT·체크리스트·형식 검증으로 정밀화

교육 4건
Khanmigo 2년 RCT, 2,379명 대학 RCT, 도쿄대 가이드 학습, NotebookLM 치과 인턴

한 문장 결론 에이전트도 학습자도 **한 번의 정답**으로 평가받지 않습니다. 다음 단계로 이어지는 수행, 즉 '지속성(carry-forward)'을 설계하고 측정하는 조직이 2027년의 생산성을 가져갑니다.

Executive Summary: Proof of Performance

Six agent papers and seven organizational or educational trials converge: what matters is not a single correct answer but sustained delivery.

Technology: tasks do not compose into projects

SlackBench re-enacted real research projects into 140 queries; across 33 model-harness pairs the best system reached about 80% and most fell below 66%. DeliverBench chained 150 subtasks into 11 professional projects over 2-15 rounds of new materials and client feedback; the best of seven systems scored 65.0 out of 100 and passed only half of subtasks at a 90-point threshold. Multi-step SWE found an 18.5% top solve rate on 69 long-horizon tasks, with 38 unsolved by any model. Capability is measured per task; value is created per project.

The fix: edit the harness, not just the answer

Flex-RSI shows recursive self-improvement pays most when an agent rewrites its own harness — prompts, tools, memory and control loop — rather than only its output. ServerBench had coding agents design complete LLM serving systems and found Claude Sonnet 5.5 delivering 2.97 dollars of latency-qualified token value per GPU dollar on Gemma E4B, 4.87x the vLLM reference, with 22 of 33 runs producing verified systems. Process design, not model choice alone, moved the metric.

Organization: deployment configuration beats adoption

Korean and international survey studies agree. Generative AI value depends on usage type (text versus multimodal) and on the mediating capabilities it builds. AI integration reaches agility only through collaboration (beta 0.406) and process flexibility (0.375), and agility predicts operational performance (0.190) — with a small total indirect effect (0.037). Across 770 Spanish firms, performance follows the interaction of deployment depth and breadth, not adoption itself.

Learning: access is free; participation is not

A two-year cluster RCT in 18 Tennessee middle schools raised math achievement by 1.3 percentile ranks per term, yet the median student messaged the tutor on only a third of practice days. A 2,379-student university trial found course-integrated tutoring reduced final grades by 0.37 SD and LMS participation by 0.90 SD. A half-day Tokyo RCT found guided questioning produced 8.1 versus 5.3 percentage points of gain. Design decides whether AI complements or displaces learning.

80% vs 66%
SlackBench best system vs most agents

4.87x
ServerBench token value vs vLLM reference

0.08
MAE with cached-trace evaluation (from 0.13)

0.037
total indirect effect, AI integration to operational performance

Audit the harness

List prompts, tools, memory and verification steps for your top agent workflow. Gains hide in the harness, not the weights.

Measure participation

Replace “licences issued” with weekly sessions and help-seeking rate on errors.

Add traceability

Require an evidence link and an explicit assumption at every checkpoint of multi-step work.

Bottom line Build for carry-forward: agents that keep traceability across steps, teams that redesign collaboration rather than buy tools, and tutors that structure participation. Measure the second step, not the first answer.

에이전트는 답을 고치기보다 '일하는 방식'을 고칠 때 강해진다

Flex-RSI의 하네스 자기수정, SlackBench의 다자 협업, ServerBench의 서빙 설계 — 세 논문이 같은 곳을 가리킨다

I 기술 (Tech) TECHNOLOGY & AGENTS

Flex-RSI (arXiv 2611.08946, 11-06) — 하네스를 고치는 자기개선

기존 벤치마크는 에이전트의 하네스(프롬프트·도구·메모리·제어 루프)를 고정한 채 산출물만 채점합니다. Flex-RSI는 매 라운드 실행기(executor)가 해법을 개선하고, 개선기(improver)가 하네스 자체를 수정하도록 했습니다. 검증 데이터에서 점수가 오를 때만 해법을 유지하고, 마지막에는 숨은 테스트로 채점합니다. 결과는 세 가지입니다. ① 자기개선은 에이전트가 **자기 해법을 직접 만든 과업**에서 가장 크게 작동하고, ② 성공·실패만 기록하는 피드백은 오픈웨이트 모델의 개선을 막으며, ③ 물리적 안정성·정확성 검사가 엄격한 과업에서 오픈웨이트 모델이 가장 크게 뒤처집니다. 하네스를 편집한 쌍(pair) 실행에서 성능 향상이 더 컸습니다.

이번 주 공개 벤치마크의 성능 지형

높을수록 좋음 · 단위 %(점수 기준 상이)



출처: SlackBench(2611.05512), DeliverBench(2611.03412), Multi-step SWE(2611.07166), TB-MathModel(2611.08237).

SlackBench — 다자 협업의 정보는 대화·코드·결정에 흩어져 있다

실제 연구 프로젝트를 프라이버시 보존 방식으로 재현해 140개 질의를 만들었습니다. 단순 검색이 아니라 어떤 대화가 중요한지 판단하고, 스레드를 따라가고, 저장소 상태를 확인하고, 상충하는 증거를 조정해야 합니다. 답할 수 없는 질문에서 **답하지 않는 능력(abstention)**도 채점합니다. 33개 모델·하네스 중 최고 약 80%, 다수가 66% 미만 — 협업 맥락은 아직 미개척지입니다.

ServerBench — 에이전트가 자기 서빙 시스템을 설계한다

목표 함수는 'GPU 비용 1달러당 지연 보정 토큰 가치'. 상태 보존 리플레이로 세션 의존성을 유지하고 독립 검증으로 제출 시스템을 확인합니다. 11개 코딩 에이전트·3개 배포 환경에서 Claude Sonnet 5.5가 Gemma E4B에서 2.97달러/달러 (vLLM 참조의 4.87배)를 기록했고, 33회 중 22회가 검증된 시스템을, 13회가 최강 베이스라인을 넘었습니다. 고성능 설계는 **용량 개선과 부하 선택을 함께 조정**했습니다.

발견 ① 하네스 편집

같은 모델이라도 하네스를 고친 쌍 실행에서 성능이 더 올랐습니다. 제품 자산은 프롬프트·도구·메모리·검증 루프입니다.

발견 ② 피드백의 질

성공·실패만 기록하면 오픈웨이트 모델은 개선되지 않습니다. 이유가 담긴 피드백이 학습 신호가 됩니다.

발견 ③ 검증의 위치

물리적 안정성·정확성 검사가 엄격한 과업에서 격차가 가장 큽니다. 배포 전 게이트가 모델 선택보다 중요합니다.

4.87× 22/33 140 33

판단 에이전트 성능의 다음 격차는 모델 가중치가 아니라 **하네스 설계**에서 생깁니다. 실무적으로는 프롬프트·도구·메모리·검증 루프를 제품의 핵심 자산으로 관리하고, '성공/실패'가 아닌 **이유가 담긴 피드백**을 로그로 남겨야 합니다.

과업의 합은 프로젝트가 되지 않는다 — 이어짐이 성능이다

DeliverBench 65.0점, Multi-step SWE 18.5%, TB-MathModel 624개 Lean 증명 — 세 벤치마크가 '지속성'을 측정한다

I 기술 (Tech) LONG-HORIZON & VERIFICATION

DeliverBench — 브리프에서 최종 납품까지

재무 실사, 법률 계약서 작성, 의약품 안전성 보고 등 11개 프로젝트를 13명의 도메인 전문가와 함께 실제 자료로 재구성했습니다. 150개 하위과업이 2~15라운드의 새 자료와 고객 피드백으로 연결되고, 참조 산출물을 주입하지 않기 때문에 **에이전트가 자기 오류를 다음 단계로 물려받습니다**. 7개 모델·하네스 가운데 최고가 최종 납품 65.0점, 90점 기준으로 하위과업의 50.0%만 통과했습니다. 한 감사 프로젝트에서는 7개 시스템 모두 지적사항을 다뤘지만 **추적성(traceability)을 보존한 시스템은 하나**뿐이었습니다.

Multi-step SWE — 자기 코드 위에 계속 쌓기

69개 과업이 각각 5~15단계의 의존적 요청으로 이어집니다. 각 단계는 앞 단계에서 에이전트가 만든 코드에서 시작하므로 초기 설계 결정과 오류가 누적됩니다. 4개 공학 영역·9개 언어, 신규 구축과 기존 오픈소스 모두 포함합니다. 9개 프론티어 모델 중 최강의 해결률은 **18.5%**, 69개 중 **38개는 어떤 모델도 해결하지 못했습니다**.

TB-MathModel은 수학적 모델링을 터미널형 과업으로 바꿨습니다. 18개 실제 CUMCM·MCM 문제에 방향 인식 단측 채점과 2층 타당성 검증(6개 불변식 + **624개 Lean 4 증명 의무**)을 붙였습니다. 최강 모델의 다중 참조 O-Eval은 64.58%, 2.54달러짜리 오픈 모델이 1/100 비용으로 59.81%를 기록했습니다.

실패의 형태 1 · 추적성 붕괴
지적은 모두 반영했지만 근거를 남기지 않으면 감사·인수인계에서 실패로 계산된다.

실패의 형태 2 · 오류 상속
앞 단계의 설계 결정이 다음 단계의 제약이 된다. 초기 구조를 검증하는 비용이 후반 수정 비용보다 싸다.

실패의 형태 3 · 검증 없는 최적화
최적화가 만든 실행 경로를 검증하지 않으면, 성능은 좋아도 시스템은 신뢰를 잃는다.

설계 원칙

각 단계의 입력·출력·검증을 문서화하고, 이전 단계 산출물을 다음 단계의 명시적 입력으로 고정합니다.

측정 지표

해결률보다 '3단계 이후 일관성'과 '추적성 보존율'을 봅니다.

비용 구조

초기 구조 검증에 투자하면 후반 수정 비용이 줄어듭니다. 18.5%의 나머지 81.5%는 대부분 초기 오류의 상속입니다.

65.0	50.0%	38/69	624
DeliverBench 최고 점수	90점 기준 통과 하위과업	어떤 모델도 못 푼 과업	Lean 4 증명 의무

한 줄 요약 과업의 합은 프로젝트가 되지 않습니다. 이어짐을 설계하고, 이어짐을 측정하십시오.

실무 적용 에이전트 도입의 성공 지표를 '첫 산출물 품질'에서 **'3단계 이후에도 유지되는 일관성'**으로 바꿔야 합니다. 체크포인트마다 근거 링크·변경 이력·가정을 남기는 것이 프로젝트형 과업의 최소 사양입니다.

도입은 사건이고, 성과는 경로다

생성형 AI의 활용 유형이 성과 경로를 가르고, AI 통합은 협업·프로세스 유연성을 거쳐야 민첩성이 된다

II 경영 (Management) ORGANIZATION & STRATEGY

생성형 AI 활용 유형 → 비즈니스 분석역량 → 조직성과 (한국빅데이터학회, 11-01)

생성형 AI 활용을 **텍스트 기반**과 **멀티모달 기반**으로 나누고, 아이디어 창출역량·정보처리역량· 인지된 데이터 과부하 감소·분석가용시간의 네 매개를 거쳐 비즈니스 분석역량과 조직성과로 이어지는 구조를 구조방정식으로 검증했습니다. 결과는 분업적입니다. 텍스트 기반은 **정보처리역량 향상과 데이터 과부하 완화**에 유의했고, 멀티모달은 **아이디어 창출역량**에 더 큰 영향을 미쳤습니다. 두 유형 모두 분석가용시간을 늘렸고, 이 요인들이 분석역량을 통해 조직성과로 이어졌습니다. "AI를 도입했는가"가 아니라 **"어떤 유형을 어떤 업무에 붙였는가"**가 성과를 설명합니다.

AI 통합 → 협업·프로세스 유연성 → 민첩성 (231명 관리자)

5,000회 부트스트랩 PLS-SEM 결과, AI 통합은 협업(β=0.406, p<.001)과 프로세스 유연성(β=0.375, p<.001)을 유의하게 높였고, 두 변수는 민첩성에 거의 동등하게 기여했습니다(β=0.245, 0.255). 민첩성은 운영성과를 예측했습니다(β=0.190, p=.003). 두 연쇄 경로 모두 유의했지만 **총 간접효과는 0.037로 작았습니다**. 저자들은 "민첩성은 AI 도입의 자동적 산물이 아니라 협업 패턴과 프로세스 구조의 실질적 재편을 필요로 한다"고 정리합니다.



β는 표준화 계수. 효과 크기는 작지만 방향과 경로가 일관된다.

텍스트형 배치

문서 요약·데이터 정리·보고서 초안 — 정보처리역량과 과부하 완화에 직접 기여합니다.

멀티모달 배치

이미지·영상·발표자료 — 아이디어 창출역량에 더 큰 효과를 냅니다.

공통 효과

분석가용시간 증가. 그 시간을 무엇에 쓸지 설계하지 않으면 성과로 이어지지 않습니다.

0.406
협업 경로 β

0.375
프로세스 유연성 β

0.190
민첩성→운영성과 β

0.037
총 간접효과

한 줄 요약 AI는 도구가 아니라 배치입니다. 무엇을 어디에, 얼마나 깊고 넓게 심었는지가 성과를 결정합니다.

해석 두 연구를 합치면 실행 순서가 나옵니다. ① 업무를 '텍스트형'과 '멀티모달형'으로 분류해 도구를 배치하고, ② 도구 도입과 동시에 협업 규칙(회의·문서·리뷰)을 재설계해야 합니다. 도구만 사면 β는 0에 가깝게 남습니다.

배치의 깊이 × 넓이, 그리고 숙련이 성과가 되는 경로

770개 스페인 기업, 600+276명 관리자, 305명 실무자 — 세 연구가 '어디에 AI를 심는가'를 측정한다

II 경영 (Management) DEPLOYMENT & CAPABILITY

배치의 깊이×넓이 (770개 대기업, 스페인)

AI 역량을 두 차원으로 재정의합니다. **깊이**는 구현의 기술적 다양성, **넓이**는 조직 내 확산 범위입니다. AI 숙련 인적자본은 깊이를, 디지털 인프라는 넓이를, 데이터 기반 문화는 둘 다를 밀어올립니다. 성과는 깊이와 넓이의 **상호작용**과 정(+)의 관계를 보였습니다 — 보완재 논리입니다. 이 연구는 AI 생산성 역설을 '도입 여부'가 아니라 **배치 구성(configuration)**으로 설명합니다.

팀 학습과 팀 창의성 (600명 + 276명, Information & Management)

연구 1은 팀 학습이 AI 역량 개발에 기여하는 미시적 기반을, 연구 2는 팀 창의성이 AI 역량을 서비스 혁신과 성과로 전환하는 가치 실현 메커니즘을 검증했습니다. 결론은 "AI 역량은 사는 것이 아니라 **팀이 배우며 만드는 것**"입니다. 조직 단위의 AI 투자와 팀 단위의 학습·창의성이 분리되면 성과 전환이 일어나지 않습니다.

AI 숙련도는 왜 곧바로 성과가 되지 않는가 (305명, 한국고용정보원 2025 자료)

IT·디자인·사무직 305명을 분석한 결과, AI 활용 숙련도가 지각된 업무성과에 미치는 **직접효과는 매개변수를 통제하면 유의하지 않았습니다**. 대신 '숙련도 → AI 활용 비중 → 업무수행 자율성 변화 → 지각된 성과'의 순차 매개가 유의했고, 두 매개를 모두 거치는 간접효과가 세 경로 중 가장 작았습니다. 해석은 명확합니다 — **교육으로 숙련을 만들어도, 실제 업무에 쓸 비중과 재량이 없으면 성과가 되지 않습니다.** 조직은 AI 교육훈련을 넘어 '적용 환경'과 '재량 설계'를 함께 손봐야 합니다.

깊이
기술적 다양성 — 한 업무에 여러 AI를 조합하는 능력. 숙련 인력이 결정한다.

넓이
조직 확산 — 부서 간 동일 도구·데이터 규칙. 인프라와 문화가 결정한다.

재량
숙련을 성과로 바꾸는 마지막 스위치. 교육보다 조직 설계의 문제다.

진단 ① 넓이

AI 도구를 쓰는 부서가 2개 이하인가? — 확산이 멈춘 조직은 성과가 나지 않습니다.

진단 ② 깊이

한 업무에 도구를 2개 이상 조합하는가? — 단일 챗봇 사용은 깊이가 아닙니다.

진단 ③ 재량

숙련자가 AI 결과를 수정·재설계할 권한이 있는가? — 없으면 숙련은 성과가 되지 않습니다.

770
스페인 대기업 표본

600+276
팀 학습·창의성 표본

305
숙련도 연구 표본

3편
국내외 조직 실증

실행 AI 전환 예산의 배분을 '교육 40 : 도구 40 : 프로세스 재설계 20'에서 '**교육 25 : 도구 35 : 프로세스·권한 설계 40**'으로 옮길 근거가 여기에 있습니다.

측정의 정밀도가 예산을 바꾼다

캐시된 실행 흔적, 항목별 체크리스트, 형식 검증 — 평가 비용을 낮추는 세 가지 방법

III 과학 (Science) MEASUREMENT & DATA SCIENCE

캐시된 흔적만으로 시스템 성능 예측 (arXiv 2611.06803)

에이전트는 실행마다 풍부한 흔적(trace)을 남깁니다. 연구는 이 흔적을 저차원 표현으로 압축해 참조 시스템들 사이의 위치로 새 시스템의 점수를 예측합니다. 문항반응이론(IRT) 모형에 이 기하학적 예측을 더하자 평균절대오차가 **SWE-bench Verified에서 0.13→0.08**(Wilcoxon $p<10^{-4}$), **Terminal-Bench 2.0에서 0.14→0.12**($p<10^{-3}$)로 개선됐습니다. 무작위 과업 **단 하나**로도 유효했습니다. 전체 벤치마크를 돌리는 비용 없이 "이 시스템이 어디쯤일지"를 추정할 수 있다는 뜻입니다.

항목별 체크리스트로 감점을 먼저 계산한다 (WeGenEditBench, 11-02)

이미지 편집 평가는 평균 점수가 지역적 실패를 가립니다. 2,740개 인스턴스·148개 과업 유형에 대해 모델 간 공유되는 사전 생성 체크리스트로 요청을 원자적 요구사항으로 분해하고, **항목별 감점과 시각적 근거를 먼저** 만든 뒤 점수로 요약합니다(편집 실행·보존·시각 품질·과업 정확성). 15개 구성 중 GPT-Image-2가 단일 이미지 편집 4.076, Seedream-5.0-Pro가 텍스트 중심 편집 4.291로 앞섰습니다. 실패를 '어디서' 일어났는지 말해주는 평가가 개선 루프(WeGenEvoStudio)의 입력이 됩니다.

형식 검증: 624개 Lean 4 증명 의무 (TB-MathModel)

수학적 모델링 답안은 '참조와 가깝고 동시에 물리적으로 실현 가능'해야 합니다. 6개 결정가능 불변식과 624개 Lean 4 증명 의무를 통과하지 못하면 점수를 0으로 만드는 하드 게이트를 두었습니다. **검증 가능한 실패**가 벤치마크의 가치를 만듭니다.

평가의 세 층위

① **과업** — 정답 여부(싸고 빠르지만 프로젝트를 설명하지 못함) ② **프로젝트** — 추적성·일관성(비싸지만 의사결정에 직결) ③ **시스템** — 비용·지연·검증(운영 가능성) 세 층위를 섞으면 "성능은 좋은데 도입은 실패"하는 착시가 생깁니다.

① 빠른 추정

IRT+흔적 예측으로 후보 20개를 3개로 줄입니다. 무작위 과업 1개로도 유효합니다.

② 프로젝트 검증

DeliverBench형 2~15라운드 과업으로 추적성·일관성을 봅니다.

③ 형식 검증

Lean·불변식 게이트로 배포 전 실패를 차단합니다.

0.08	2,740	148	624
MAE (개선 후)	WeGenEditBench 인스턴스	과업 유형	Lean 4 증명 의무

한 줄 요약 측정 비용을 낮춘 팀이 실험을 더 많이 하고, 실험을 더 많이 한 팀이 더 빨리 배웁니다.

판단 평가 예산은 곧 실험 횟수입니다. IRT 기반 추정으로 후보를 빠르게 걸러내고, 프로젝트형 벤치마크로 최종 후보만 검증하며, 형식 검증으로 배포 전 실패를 걸러내는 **3단 파이프라인**이 현실적인 표준이 됩니다.

같은 AI 튜터, 정반대의 결과 — 무엇이 달랐나

18개 중학교 2년 RCT는 작은 플러스, 2,379명 대학 RCT는 0.37SD 마이너스. 차이는 '의무 시간'과 '대체 여부'였다

IV 교육 (EduTech) LEARNING OUTCOMES

연구	설계	결과와 해석
One Click Away: Khanmigo **18개 중학교 · 2년 · 5학기**	학군 단위 군집 무작위배정. 기존 보충수업(25~40분) 시간에 의무 배치. 코칭형(정답 차단) 설정.	학기당 **+1.26 백분위**(SE 0.60), 통제집단 0.05SD, 전체 모집단 0.039SD. 2년차 완전 참여 시 0.14SD. 그러나 학생 96%가 1회 이상 사용했어도 중앙값 기준 연습일의 1/3에만 메시지, 오답 세션의 17%에서만 도움 요청. 메시지는 대부분 단답·추천 프롬프트 클릭. **AI 없는 Khan Academy 연습 효과와 유사**.
교과 통합 생성형 AI 튜터 **2,379명 · 30개 강좌**	동일 교과 섹션 단위 무작위배정. 수업에 통합된 튜터 접근 제공.	최종 성적 **-0.37SD**, LMS 참여 **-0.90SD**. 첫 세대 학생의 손실이 더 컸음. "AI 튜터가 기존 학습 활동을 밀어내면서 성과를 만들지 못했다" — 대체(displacement) 효과.

두 연구를 가르는 세 가지 조건

① **시간의 소유** — Khanmigo는 이미 존재하는 보충수업 시간에 들어가 추가 시간을 만들었습니다. 대학 튜터는 학생의 기존 학습 시간을 나눠 썼습니다.
② **답의 차단** — 코칭형은 정답을 주지 않고 질문으로 유도합니다. 자유형은 즉답이 가능해 연습을 생략하게 만듭니다.
③ **측정 대상** — 시험 점수만 보면 참여의 질을 놓칩니다. 메시지 빈도·오답 시 개입률이 실제 효과의 선행지표입니다.

숫자로 보는 참여

96% — 최소 1회 사용
1/3 — 중앙값 학생이 연습일에 메시지를 보낸 비율
17% — 오답이 발생한 세션에서 도움을 요청한 비율
0.14SD — 2년차 '완전 참여' 시 환산 효과

왜 대학에서는 역효과였나

① 기존 학습 시간을 대체 ② 즉답 가능 ③ 참여 강제 없음 — 세 조건이 겹치면 성적이 내려갑니다.

왜 중학교에서는 플러스였나

① 의무 시간에 배치 ② 정답 차단 ③ 2년 누적 참여 — 효과는 참여의 함수입니다.

제품 사양으로 옮기면

정답 차단·오답 개입·주간 세션 리포트를 기본 탑재하고, 도입 KPI를 참여로 정의합니다.

+1.26
백분위/학기

-0.37SD
최종 성적

-0.90SD
LMS 참여

96%
최소 1회 사용률

시사점 "AI를 제공했다"는 도입 지표이고, "대화가 일어났다"는 성과 지표입니다. 학교·기업 교육 담당자는 라이선스 수가 아니라 **주간 대화 세션 수와 오답 개입률**을 계약의 KPI로 넣어야 합니다.

질문을 유도하면 1.5배, 자유롭게 두면 주의가 흩어진다

도쿄대 300명 반나절 RCT와 치과 인턴 NotebookLM 연구 — 설계와 매체가 학습을 가른다

IV 교육 (EduTech) GUIDED LEARNING & MULTIMODAL

Guided Learning (Google, 11-03) — 도쿄대 300명

반나절 동안 통계 수업을 혼자 공부하게 하고, 절반에게 Gemini 앱의 가이드 학습을 제공했습니다. 두 집단 모두 온라인 자료를 쓸 수 있었고, 가이드 집단은 정답 대신 단계별 질문·연습 퀴즈·맞춤 시각 자료를 활용했습니다. 사전-사후 점수 변화는 일반 도구 **+5.3%p**, 가이드 학습 **+8.1%p**로 **1.5배**의 학습 이득을 보였습니다. 반나절이라는 짧은 세션에서 나온 차이라는 점이 중요합니다 — 설계만 바꿔도 효과가 납니다.

NotebookLM 강의 vs 대면 강의 (BMC Medical Education, 11-05)

치과 인턴을 무작위 배정해 동일 교재로 AI 생성 영상 강의와 대면 강의를 비교했습니다. 진단(p=0.327)·생활치수 치료(p=0.411)에서는 유의차가 없었고, 근흡수 문항에서만 AI 집단이 **87.8% vs 83.0%**(p=0.029)로 앞섰습니다. 만족도·논리적 흐름·임상 적용 자신감은 AI 집단이 높았지만, **자기보고 주의 분산은 74.5% vs 38.3%**(p=0.003)로 AI 집단이 훨씬 높았습니다. "편리하지만 주의가 흩어진다"는 것이 멀티모달 강의의 실제 프로필입니다.

설계 원칙 1
정답을 주지 않고 질문으로 유도한다. 즉답형은 연습을 생략시킨다.

설계 원칙 2
짧은 세션에서도 '단계별 확인'을 넣는다. 반나절만으로 1.5배 차이가 난다.

설계 원칙 3
멀티모달은 만족도를 올리지만 주의 분산을 함께 올린다. 길이·속도·구간 확인이 필요하다.

강의형 AI

전달 효율↑, 만족도↑, 주의 분산↑ — 구간 확인과 길이 통제가 필요합니다.

튜터형 AI

상호작용↑, 참여 미보장 — 오답 개입과 질문 유도가 필요합니다.

혼합 설계

10분 강의 → 2문항 확인 → 1회 질문 유도 → 요약의 4단 구조가 현실적입니다.

+8.1%p
가이드 학습

+5.3%p
일반 도구

87.8%
근흡수 문항 AI

74.5%
주의 분산 자기보고

한 줄 요약 전달과 연습은 다른 제품입니다. 강의는 주의를 관리하고, 튜터는 참여를 설계해야 합니다.

해석 'AI 강의'와 'AI 튜터'는 다른 제품입니다. 강의는 전달 효율을 높이지만 주의를 잃기 쉽고, 튜터는 상호작용을 만들지만 참여를 강제하지 못합니다. 제품 정의 단계에서 **전달(delivery)과 연습(practice)을 분리**하고 각각의 지표를 세워야 합니다.

세 개의 축: 지속성 · 참여 · 측정

기술·조직·교육 연구가 같은 결론으로 수렴하는 지점 — 한 번의 정답이 아니라 이어지는 수행

V 교차 분석 CROSS-DOMAIN SYNTHESIS

축	기술 (에이전트)	조직 (경영)	교육 (학습)
지속성 **Carry-forward**	DeliverBench 65.0점, 추적성 보존 1/7. Multi-step SWE 18.5%, 38/69 미해결. 앞 단계 오류가 다음 단계의 제약이 된다.	깊이×넓이의 상호작용(770개 기업). 도구만 사면 배치가 얕고 좁아 성과가 나지 않는다.	2년 RCT의 효과가 2년차에 집중 — 참여가 성숙해야 성과가 나타난다.
참여 **Participation**	SlackBench 66% 미만 다수. 정보가 흩어진 환경에서 '무엇을 볼지' 정하는 능력이 병목.	AI 숙련도 → 활용 비중 → 자율성의 순차 매개. 재량이 없으면 숙련은 성과가 아니다.	Khanmigo 중앙값 1/3 연습일, 오답 세션 17%. 대학 RCT는 참여가 대체되어 -0.37SD.
측정 **Measurement**	IRT+흔적 예측으로 MAE 0.13→0.08. 항목별 체크리스트, 624개 Lean 증명.	β=0.406·0.375·0.190의 경로, 총 간접효과 0.037 — 작지만 방향이 분명한 효과.	참여 로그·오답 개입률이 선행지표. 시험 점수만 보면 설계 결함을 놓친다.

세 축을 하나로: '이어지는 수행'의 정의

에이전트에게는 **추적성**, 조직에게는 **배치의 상호작용**, 학습자에게는 **주간 대화 세션**이 같은 개념의 다른 이름입니다. 모두 "다음 단계에서도 유지되는가"를 묻습니다. 이 질문을 제품·평가·교육 설계의 공통 언어로 삼으면, 부서마다 다른 KPI를 하나의 지표 체계로 묶을 수 있습니다.

반대 증거도 기록한다

모든 연구가 AI의 효과를 지지하지는 않습니다. 대학 RCT의 음(-)의 효과, Khanmigo의 작은 효과 크기, 조직 연구의 작은 간접효과(0.037)는 "AI는 조건부로만 작동한다"는 증거입니다. **조건을 설계하지 않으면 평균은 0에 가깝습니다.**

한 줄 요약 기술·조직·교육의 KPI는 결국 하나의 질문으로 모입니다 — 다음 단계에서도 유지되는가.

3	65.0	0.037	1/3
수렴 축 수	DeliverBench 최고점	조직 총 간접효과	Khanmigo 참여일 비율

통합 지표 예시 — 부서별 KPI를 하나의 표로

문서 1건당 근거 링크 수 · 주간 AI 세션 수 · 오답 개입률 · 배치 깊이·넓이 지수 · 추적성 보존율 · 재작업률 — 여섯 지표를 제품·평가·교육에 공통 적용하면 "도입했는데 성과가 없다"는 논쟁을 데이터로 끝낼 수 있습니다.

실행 프레임워크: 진단 → 설계 → 측정 → 확장

논문이 말한 조건을 90일 실행으로 옮기는 네 단계

VI 실행 프레임워크 PRACTICAL FRAMEWORK

01 지속성 진단 (2주)
핵심 업무를 5~15단계로 분해하고, 각 단계의 입력·검증·인수인계 문서를 확인한다. 추적성이 끊기는 지점을 표시한다.

02 참여 설계 (4주)
이미 존재하는 시간(수업·회의·리뷰)에 AI를 넣는다. 새 시간을 만들지 않는다. 정답 차단·질문 유도·오답 개입을 기본값으로 둔다.

03 측정 체계 (지속)
1차 지표는 참여(주간 세션·오답 개입률), 2차는 산출물 품질, 3차는 비용·시간. 시험 점수는 마지막에 본다.

04 확장과 거버넌스 (분기)
배치의 깊이(도구 다양성)와 넓이(부서 확산)를 함께 넓힌다. 권한·감사·개인정보 처리 문서를 제품 사양에 포함한다.

각 단계의 산출물
① 과업 분해도·끊김 지점 목록
② 시간 배치표·개입 규칙서
③ 지표 정의서·대시보드
④ 배치 지도·거버넌스 문서

실패 신호 체크리스트
□ 파일럿이 '새 시간'을 요구한다
□ 성과 지표가 시험 점수·만족도뿐이다
□ 근거 링크 없는 산출물이 승인된다
□ 도구는 늘었는데 협업 규칙은 그대로다
□ 개인정보 처리 문서가 제품 사양에 없다

한 줄 요약 진단 없는 도구 도입은 예산만 늘리고, 측정 없는 확장은 학습을 줄입니다.

2주	4주	90일	분기
진단	참여 설계	측정 안정화	확장·거버넌스

90일 일정표
1~2주 과업 분해·끊김 지점 목록 → 3~4주 파일럿 2개 선정과 시간 배치표 → 5~8주 지표 수집·주간 리뷰 → 9~12주 리뷰 결과로 확장 여부 결정. 확장은 측정 체계가 안정된 뒤에만.

주의 네 단계는 순서가 중요합니다. 측정 체계 없이 확장하면 예산은 늘고 학습은 줄어듭니다. 반대로 진단 없이 도구를 고르면 'AI 없는 Khan Academy와 같은 효과'에 머뭅니다.

시사점: 다음 회계연도의 경쟁력은 '이어짐'에서 갈린다

연구 아홉 편과 실증 네 건이 남긴 다섯 개의 명제 — 창업가·경영진·교육기관을 위한 권고

VII 전략적 시사점 STRATEGIC TAKEAWAYS

1. 에이전트 제품은 '하네스'가 경쟁력이다
프롬프트·도구·메모리·검증 루프를 자산으로 관리하고, 이유가 담긴 피드백 로그를 남겨라. 모델 교체보다 하네스 개선이 성능을 더 올린다.

2. 프로젝트형 과업에는 '추적성'을 기본값으로
체크포인트마다 근거·가정·변경 이력을 남기지 않으면, 결과가 맞아도 감사·인수인계에서 실패한다.

3. 조직은 '깊이×넓이×재량'을 함께 설계
교육 25 : 도구 35 : 프로세스·권한 40. 숙련이 활용 비중과 자율성으로 이어지지 않으면 성과는 0에 가깝다.

4. 학습 제품은 '참여'를 사양으로 판다
주간 대화 세션·오답 개입률을 계약 KPI로. 정답 차단·질문 유도가 기본값이고, 새 시간을 만들지 말고 기존 시간에 넣어라.

5. 평가는 3단 파이프라인으로
IRT 기반 빠른 추정 → 프로젝트형 검증 → 형식·물리 검증. 평가 비용을 낮춘 팀이 실험 횟수를 늘린다.

콘데리 관점
SkillBridge의 '실무 과업 수행 기록'은 이 논문들이 말한 추적성·참여·측정의 세 축을 하나의 인증 구조로 묶는 시도다. 지방 학교·중소기업에서 먼저 검증한다.

참고문헌 (References)

Flex-RSI (arXiv 2611.08946) · SlackBench (2611.05512) · ServerBench (2611.06841) · DeliverBench (2611.03412) · Multi-step SWE (2611.07166) · TB-MathModel (2611.08237) · Task-Efficient Agent Evaluation Using Cached Traces (2611.06803) · WeGenEditBench (2026-11-02) · 생성형 AI의 활용이 비즈니스 분석역량과 조직성과에 미치는 영향(한국빅데이터학회, 2026-11) · AI 통합이 조직 민첩성에 미치는 영향(한국컴퓨터정보학회논문지 31(8), 2026) · AI capability and organizational performance(Information & Management, 2026-10-28) · Beyond AI Adoption (JAIS/PACIS 2026) · The Effect of AI Skill Level on Perceived Work Performance(Journal of Skills and Qualifications 15(3), 2026) · One Click Away: AI Tutoring with Khanmigo (EdWorkingPaper 26-1551) · Course-Integrated AI Tutoring RCT (EdWorkingPaper 26-1598) · Google for Education, Guided Learning RCT (2026-11-03) · BMC Medical Education (2026-11-05)

한 줄 요약 능력은 측정되고, 성과는 이어집니다. 이어짐을 설계하는 조직이 다음 회계연도의 원가를 결정합니다.

동시 발행 · 시장 인사이트
동의의 병목: AI 인프라가 표를 얻는 값 (CONDERI-IR-2026-11-10)

동시 발행 · 포커스 칼럼 4호
허가의 경제학: AI 인프라의 다음 병목은 전기가 아니다

동시 발행 · 글로벌 옴니버스 2호
6대 매체 교차 분석 + 전문 차트 5종 (CONDERI-GO-2026-11-10)

고지 본 리포트는 공개된 논문·실증 자료를 콘데리 연구실이 해석한 것이며 투자·법률 자문이 아닙니다. 수치는 각 논문이 보고한 값이고, 효과 크기의 일반화에는 표본·기간·맥락의 차이를 고려해야 합니다.