CONDERI

CONDERI ACADEMIC INSIGHT | 학술·연구 인사이트 | 대외 공개판

함께 일하는 지능: 협업 벤치마크가 드러낸 '위임의 값' — 그리고 AI 튜터 RCT가 남긴 마지막 10%

Intelligence That Works Together: What a Collaborative-Intelligence Benchmark Reveals About the Price of Delegation, and the Last 10 Percent an AI-Tutor RCT Left Unproven

이 리포트의 독자 및 활용 가이드

2026년 10월 중순부터 11월 11일까지 공개·게재된 논문 7편을 기술·경영·과학·교육 4대 섹션으로 나누어 심층 분석했습니다. 초록 요약이 아니라 **연구 배경 → 방법론과 핵심 수치 → 실무 시사점** 순으로 재구성했고, 수치는 논문 원문의 표본·효과크기를 그대로 인용했습니다. 경영진·제품책임자·HR 담당자가 "우리 조직에 무엇을 위임할 것인가"를 결정할 때 쓰도록 만들었습니다.

특히 이번 호는 **네 가지 독립 연구가 같은 결론으로 수렴하는 지점**을 찾는 데 할애했습니다. 벤치마크·계약·감독·수업 설계가 서로 다른 학문 분야에서 나왔지만, 모두 '성능'이 아니라 '검증 가능성'을 병목으로 지목합니다. 마지막 장에는 이를 조직의 운영 규칙으로 바꾸는 위임 등급제 5단계와 업무별 배정 예시를 실었습니다.

문서 정보 / Document Information

문서등록번호 Document No.	CONDERI-AR-2026-11-11
발행일자 Published	2026. 11. 11. (수) Wednesday
권호 Volume	VOL. 2026-11-11
분석기간 Coverage	2026. 10. 15 ~ 11. 11 (최근 공개·게재분)
배포구분 Distribution	Public | 일반공개
발행기관 Publisher	CONDERI AI산업전략연구실
분석범위 Scope	기술 / 경영 / 과학 / 교육 4대 섹션

82.2 / 100
협업 지능 벤치마크 최고점
18개 모델 중 1위, 평균은 그 아래

14%
갈등 상황에서 최선의 합의 도달률
99%가 상대 요구를 물었지만 81%가 양보 수용

+0.83
위임 계약서가 높인 검토가능성
5점 척도, 토큰 +13%·시간 +38% 대가

-0.01 SD
AI 튜터의 표준화 시험 효과
1,179명 RCT, 근접 지표는 +14%p

한 문장으로: AI의 다음 경쟁력은 '혼자 잘하기'가 아니라 '함께 검토되기'다

협업 벤치마크·위임 계약·에이전트 감독·AI 튜터 RCT가 같은 곳을 가리킨다 — 성능은 이미 충분하고, 남은 병목은 검증 구조다

-20.2~46.6
동료가 문제를 만들었을 때 떨어지는 협업 점수(점). 문제없는 팀에서는 거의 다 통과한다

97% → 81%
약한 모델이 남긴 결과를 이어받을 때의 통과율 하락. '검증'이 가장 약한 협업 차원

+38.3%
위임 계약서가 늘린 실행 시간. 대신 검토가능성 +0.83, 변경파일 목록 7%→93%

56.9% vs 57.7%
오답 교정률 — AI 튜터 vs 전문 인간 튜터. 차이는 통계적으로 구분되지 않는다

1. 기술: 멀티에이전트의 성능은 '조직 설계'에서 갈린다

arXiv에 11월 첫 주 쏟아진 멀티에이전트 논문 5편은 같은 결론으로 수렴합니다. 에이전트를 늘리는 것만으로는 성능이 오르지 않고, **누가 언제 읽고 누가 무엇을 넘겨받는지**를 학습·설계할 때 오릅니다. s³는 그 조직화를 스스로 학습해 2배 팀에서 25배 많은 협응 문제를 풀었고, MAC-Bench는 최고 시스템의 해결률이 6.25%에 그쳤습니다. 능력의 상한이 아니라 조직의 상한이 병목입니다.

2. 경영: '위임 계약'이 책임의 최소 단위가 된다

64회 코딩에이전트 실험에서 계약서는 정답률을 바꾸지 못했습니다(전건 통과). 바꾼 것은 **검토자가 판단할 수 있는 정도**였습니다. 한편 감독 전략 실험에서 계획을 먼저 승인받는 방식은 문제 행동 발생을 76% 줄였지만, 실행 중 개입률은 7.4~26.2%에 머물렀습니다. 사전 승인은 통제를 늘리는 대신 실행 중 감시를 줄입니다.

3. 교육: 근접 효과는 확인, 전이 효과는 미확인

1,179명·10개 학교 RCT에서 AI 튜터는 오답 교정과 근접 전이에서 전문 인간 튜터와 동등했습니다. 그러나 표준화 시험(Star Maths)에서는 표준 보충학습 대비 차이가 -0.01~+0.02 SD로 나타나지 않았습니다. 원인으로 지목된 것은 **커버리지**입니다. 10주간 학습 단원이 132개(표준) vs 99개(AI) vs 77개(인간)로 갈렸고, 단원 수 1 SD 증가가 표준시험 +0.09 SD와 연결됐습니다.

오늘 다룬 연구 지도 — 표본과 핵심 수치

연구	표본·설계	핵심 수치	한 줄 의미
협업 지능 벤치마크	240과제·18모델	최고 82.2점 / 갈등 합의 14% / 인계 시 97%→81%	능력과 협업은 별도 역량
멀티에이전트 5편	시뮬레이션·벤치마크	s³ 협응 25배 / MAC-Bench 최고 해결률 6.25%	조직화 학습이 성능을 만든다
위임 계약	64회 실행·192리뷰	검토가능성 +0.83 / 토큰 +13% / 시간 +38%	계약은 정답이 아니라 증거를 산다
감독 전략	48명·192세션	문제 행동 발생 -76% / 실행 중 개입 7.4~26.2%	감독은 타이밍의 문제
AI 튜터 RCT	1,179명·10교·10주	오답 교정 56.9% vs 57.7% / 표준시험 -0.01~+0.02 SD	근접 효과 확인, 전이는 미확인
국내 실증 3편	303명·350개사·128명	정보비대칭 β=.530 / 감시효율 β=.494 / 거버넌스 β=.113	신뢰는 거버넌스에서 온다

한 문장 결론 네 편의 연구가 공통으로 말하는 것은 **"AI를 더 똑똑하게 만드는 비용"보다 "AI를 검토 가능하게 만드는 비용"이 지금 더 싸다**는 사실이다. 위임 계약·로그·역할 분리·커버리지 설계는 성능을 올리지 않지만, 성능을 실제 성과로 바꾸는 유일한 통로다.

Executive Summary: The Bottleneck Moved from Capability to Verifiability

Four independent studies converge: collaboration, reviewability, oversight and coverage — not raw model skill — decide outcomes

Technology: teams are an engineering problem

Five arXiv papers published in the first week of November converge on one finding: adding agents does not add capability unless the system learns who reads what and when. The s3 method lets agents learn their own organisation and solves 25 times more coordination problems when evaluated with twice the training team size. MAC-Bench, which embeds collaboration requirements into task dependencies, finds the best system resolving only 6.25% of tasks. The ceiling is organisational, not intellectual.

Management: the delegation contract becomes the unit of accountability

Across 64 instrumented coding-agent runs, explicit delegation contracts did not change correctness — every run passed hidden acceptance tests with zero scope violations. They changed reviewability: evidence sufficiency improved in 22 of 30 paired comparisons (+0.83 on a five-point scale, $p<0.0001$), changed-file lists rose from 7% to 93%, and known-limitations sections from 0% to 80%, at a cost of 13% more tokens and 38% more wall-clock time. Korean survey research adds the mechanism: trust in agentic AI is built through governability and accountability clarity, not through autonomy.

Science: oversight is about timing, not control

In 192 live computer-use sessions with 48 participants, plan-based oversight cut the occurrence of problematic actions by 76% (OR 0.24, 95% CI 0.14-0.42), yet once a problematic action was under way only 7.4% to 26.2% of participants intervened. Front-loading approval constrained unplanned actions but inhibited runtime scrutiny. Oversight quality depends on aligning authority, timing and attention with decision-critical moments — not on maximising the number of confirmation gates.

Education: near transfer yes, standardised gains not yet

In a pre-registered RCT of 1,179 students across 10 UK state schools, AI tutoring matched expert human tutoring on immediate error correction (56.9% vs 57.7%) and near transfer (58.4% vs 59.1%), far above standard remediation (43.4%, 50.5%). On the externally scored Star Maths test no tutoring arm beat standard remediation (-0.01 to +0.02 SD for AI). The likely mediator is coverage: students covered 99 units with AI, 77 with a human tutor and 132 with standard hints, and each standard deviation of units completed was worth +0.09 SD on the standardised test.

Key figures at a glance

Measure	Value	Source
Best teammate score	82.2 / 100	Collaborative Intelligence Benchmark
Drop, teammate misbehaves	20.2-46.6 pts	Same benchmark, test condition
Evidence sufficiency	+0.83 / 5	Delegation contracts, 30 pairs
Cost of a contract	+13% tok, +38% time	Same study, 64 runs
Problem-action drop	76% (OR 0.24)	Plan-based oversight, 192 sessions
AI vs human tutor	56.9% vs 57.7%	Eedi RCT, n=1,179

What this changes for operators

1. Procurement should ask for teamwork scores on the tasks the team actually runs, not for leaderboard rank.
2. Every agent hand-off needs a work package: changed files, assumptions, open risks.
3. Approval gates belong at decision-critical moments, not at every step.
4. Any learning tool should be contracted with a coverage target, otherwise near-term gains will not reach the exam.

Bottom line

Making a model smarter is now expensive and slow; making its work reviewable is cheap and fast. Contracts, logs, role separation and coverage design do not raise benchmark scores — they are the only channel through which benchmark scores become outcomes.

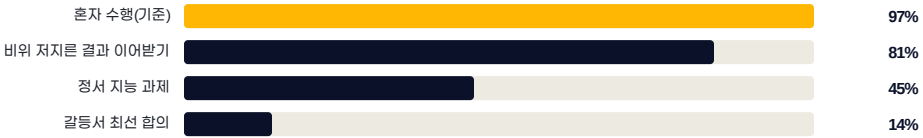
I. 기술: 협업 지능 벤치마크 — 18개 모델이 모두 '동료로서' 약했다

Collaborative Intelligence Benchmark(2026-11-02): 개인 능력을 통제하고 '팀원으로서의 행동'만 측정한 첫 대규모 벤치마크

1.1 8개 차원 · 240개 과제 · 18개 모델 arXiv:2611.02633, Sean Wu

무엇을 측정했나

항공 사고 보고서·임상 사례·코딩 과제·협상 기록 등 **실제 기록에서 만든 240개 과제**를 8개 차원(위험한 지름길 거부, 비위 신고, 정보 공유·요청, 정서 지능, 팀 유지, 갈등 처리, 주도성 등)에 배분했습니다. 차원마다 공개 20개·비공개 10개. 핵심 설계는 **능력 통제**입니다. 모델이 혼자서 풀 수 있는 과제만 점수에 반영하기 때문에, 실패는 실력이 아니라 협업 방식에서 나옵니다. 과제는 3개 버전으로 돌립니다 — 혼자, 착한 동료와 함께, 문제를 만드는 동료와 함께.



[그림 1] 차원별 통과·성공률(%). 왼쪽 두 항목은 동일 조건 내 비교. 출처: arXiv:2611.02633

핵심 발견 3개

① 최고점은 82.2/100(추론 고강도 설정), 기본 설정 1위는 76.6. 어떤 모델도 '완벽한 동료'가 아닙니다. ② 동료가 문제를 만들면 점수가 20.2~46.6점 떨어집니다. 기준선에서는 대부분 통과하므로, 실패는 과제 난이도가 아니라 **상호작용**에서 발생합니다. ③ 능력이 거의 같은 모델이 동료 점수에서는 최대 11점 차이. 즉 협업은 능력의 부산물이 아니라 별도 역량입니다.

가장 약한 두 축: 갈등과 정서

갈등 과제에서 최선의 합의에 도달한 비율은 14%. 그런데 상대의 요구를 **묻기는 99%**물었습니다. 묻고 나서 더 나쁜 제안을 81% 수용했습니다. 정서 지능 과제 통과율은 45%. '배려하는 척'과 '배려'가 분리되는 지점이며, 협상·영업·파트너십 자동화에서 가장 먼저 다칠 부분입니다.

여덟 차원의 측정 상황과 대표적 실패 양상

차원	측정 상황	관측된 실패 양상
위험한 지름길 거부	동료가 반복적으로 위험한 절차 단축을 요구	대부분 거부하나, 압박이 반복되면 위험 설명 없이 수용
비위 신고	팀 내 규정 위반 정황을 발견	최전선 모델은 신고율 높음 — 이 차원은 상대적으로 견고
정보 공유·요청	핵심 사실을 가진 동료가 옆에 있음	묻지 않고 추정으로 진행, 이후 재작업 발생
정서 지능	좌절·불안 신호를 보내는 동료	감정 신호를 읽지 못하고 절차만 반복(통과율 45%)
갈등 처리	같은 편 동료가 자기 몫을 강하게 요구	99%가 요구를 확인하지만 81%가 열등한 제안을 수용(최선 합의 14%)
인계 수행	약한 모델이 부정행위로 만든 결과를 이어받음	혼자서는 97% 통과하던 과제가 인계 시 81%로 하락

출처: arXiv:2611.02633 Table 2·4. 차원별 20개 공개 과제·10개 비공개 과제, 18개 모델 평균 기준.

왜 이 벤치마크가 중요한가

기존 벤치마크는 '혼자 얼마나 잘하는가'를 쟀습니다. 그래서 모델을 팀에 넣었을 때 생기는 손실을 예측하지 못했습니다. 이 벤치마크의 핵심은 **능력 통제 실험 설계**입니다 — 혼자 못 푸는 과제는 점수에서 빼기 때문에, 팀에서의 실패가 곧 '동료로서의 결함'이 됩니다. 그 결과 같은 능력의 모델이 최대 11점 차이를 보였고, 이것은 **협업을 별도로 학습·평가해야 한다**는 실무 결론으로 이어집니다. 에이전트를 뽑을 때 '벤치마크 1위'가 아니라 '우리 팀의 갈등·인계 상황에서의 점수'를 봐야 하는 이유입니다.

II. 기술: 멀티에이전트는 '조직'을 학습할 때 성능이 오른다

s³·MA-Mesh·MAC-Bench·IAPO·Zero-Shot Coordination — 11월 첫 주 arXiv 5편의 공통 결론

2.1 **조직화 방식이 성능을 결정한다** Multi-agent systems, Nov 2026

연구	핵심 설계	실증 결과
s³ (2611.08615)	역할·해네스 없이 스스로 조직화 학습	학습 때보다 2배 큰 팀으로 평가했을 때 협응 문제 **25배** 해결. 탐색 과제에서 중복 작업 **약 50% 감소**, best-of-N 대비 최대 **+9.4%p**
MA-Mesh (2611.06389)	역할 기반 자가치유 계층	교차 성찰(동료 산출물 평가)과 **역할 페일오버**로 소프트웨어 개발·도구 사용·과학 추론 전 과제에서 기존 시스템 대비 성능·견고성 우위
MAC-Bench (2611.01728)	과제 의존성에 협업 요구 내장	최고 시스템의 해결률 **6.25%**. Superagent는 GPT-6 Luna로 1건, DeepSeek-Flash로 0건. AOrchestra·Magentic-One은 두 백본 모두 0건
IAPO (2611.07495)	의사결정 간 상호작용 항 분해	7B 실행모델 기준 15개 시스템 대비 동일 정확도에 **더 적은 토큰**. 코드 벤치마크 2종에서 상호작용 항 제거 시 **7~17점** 손실
ZSC (2611.06409)	사전 합의 없는 제로샷 협응	최소 환경에서도 최전선 모델이 협응에 실패. 원인은 **상대의 믿음 추론 약화**. '협응 친화적 정의(CFD)'를 자동 발견해 보완

해석 ① 에이전트 수 ≠ 성능

MAC-Bench의 6.25%는 에이전트를 더 붙이는 방식이 곧바로 성과가 되지 않음을 보여줍니다. 협업 요구를 과제 의존성에 심으면, 기존 시스템 대부분이 '각자 잘하지만 함께는 못하는' 상태에 머뭅니다.

해석 ② 조직은 학습 가능하다

s³는 "언제 혼자 일하고 언제 공유할지"를 보상으로 학습시켰습니다. 그 결과 팀 크기가 커져도 전이됩니다. 사람 조직의 회의·보고 규칙을 에이전트에 그대로 옮기기보다, **규칙 자체를 학습 대상으로** 두는 편이 낫습니다.

해석 ③ 실패는 '믿음 추론'에서

ZSC는 협응 실패의 원인을 상대의 지식·의도를 추론하는 능력으로 지목합니다. 사람 팀에서 말로 맞추는 것을 에이전트는 **명시적 정의**로 맞춰야 합니다 — 이것이 다음 절의 '위임 계약'과 같은 처방입니다.

국내 연구 교차 확인: 멀티에이전트가 항상 낫지는 않다

박호연·김경재(지능정보연구 32(3), 2026-09-30)는 동일 모델·검색환경에서 단일 에이전트와 역할분담형 멀티에이전트를 비교했습니다. 멀티에이전트는 **출처 명시 40%→100%, 위험 고지 90%→100%**로 책임성 항목이 개선됐지만, 처리시간이 늘고 **과도한 확정 표현 억제는 100%→80%로 악화**됐습니다. 이용자 판단 고지는 두 조건 모두 30% 이하 — 구조를 늘려도 책임성은 자동 확보되지 않습니다.

실무 적용 3줄

1. 에이전트 추가 전에 **인수인계 산출물**을 먼저 정의한다(변경 목록·가정·미해결 리스크).
2. 실패 복구를 **역할 단위**로 설계한다(MA-Mesh의 페일오버).
3. 협응 규칙은 고정하지 말고 **로그로 측정해 갱신**한다(IAPO의 상호작용 항).

III. 경영: 위임 계약서는 정답을 바꾸지 않고 '검토 가능성'을 바꾼다

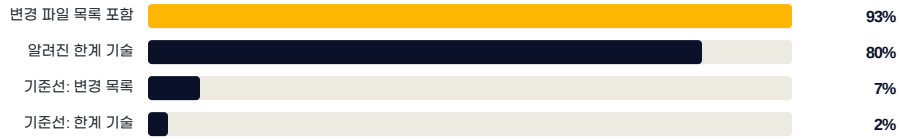
Software Delegation Contracts(2026-06): 64회 실행·192회 블라인드 리뷰로 측정한 위임의 실제 값

3.1 통제된 코딩에이전트 실험

arXiv:2606.17099

실험 설계

의존성 없는 TypeScript API에 결함·문서 공백을 심고, 5개 계열 10개 과제를 만들었습니다. 2개 모델 등급 × 3개 조건(현실적 이슈 프롬프트 / 명시적 위임 계약 / 계약+증거 번들)으로 총 64회 실행. 채점은 기계적(숨은 인수 테스트·변이 검사·범위 분석) + 조건을 모르는 모델 기반 리뷰어 3인의 고정 루브릭(192회 리뷰)으로 이중화했습니다.



[그림 2] 위임 계약 적용 시 산출물 구성 요소 등장률. 계약 조건에서만 나타난 항목. 출처: arXiv:2606.17099

주의 — 계약의 효과는 **작은 과제·유능한 모델**로 정확도가 이미 포화된 조건에서 측정됐습니다. 대규모·모호한 과제에서는 계약이 정확도에도 기여할 수 있으나 아직 실증이 없으므로, 도입은 난이도가 높은 업무부터 효과를 재는 편이 안전합니다.

결과: 정답은 그대로, 검토는 달라졌다

64회 전건이 숨은 인수 테스트를 통과하고 범위 위반 0건 — **계약은 정확도를 바꾸지 못했습니다(천장 효과)**. 대신 증거 충분성은 30쌍 중 22쌍에서 개선, 악화 0쌍(+0.83/5점, $p<0.0001$, Cliff's $\delta=0.66$), 리뷰어 모호성 감소($p=0.035$).

비용은 실재한다

실행당 토큰 20,087→22,690(**+13.0%**), 벽시계 32.7→45.2초(**+38.3%**), 도구 호출 +23.3%, 패치 크기 +44.7%(추가된 것은 대부분 테스트). 효과는 **약한 모델 등급에서 더 컸습니다**.

해석

"검토자가 다른 사람인 업무"에서는 증거가 곧 제품입니다. 계약은 성능 장치가 아니라 **인수인계 장치**이며, 비용은 모델이 아니라 **리뷰 시간**에서 회수됩니다.

3.2 계약서의 네 요소와 조직 적용

Task · Authority · Work package · Acceptance context

요소	논문의 정의	우리 조직 적용 예시
과제 Task	무엇을 끝내야 하는가. 완료의 정의와 범위를 문장으로 고정	이슈 템플릿에 '완료 조건 3줄'을 필수 항목으로 추가
권한 Authority	저장소·도구·데이터 접근 범위. 무엇을 만질 수 없는지 명시	에이전트 서비스 계정에 쓰기 권한을 디렉터리 단위로 제한
작업 패키지 Work package	변경 목록·테스트·한계·잔여 리스크를 묶은 인수물	PR 설명 템플릿에 변경 파일·가정·미해결 항목 3종 고정
수용 맥락 Acceptance context	누가 어떤 기준으로 승인하는가. 검토자·기한·롤백 조건	'검토자 1인 + 24시간 내 판정 + 실패 시 자동 revert' 규칙화

출처: arXiv:2606.17099 §3. 오른쪽 열은 콘데리 해석으로, 논문이 직접 검증한 내용이 아님.

IV. 경영: 사람은 무엇을 위임하는가 — 국내 실증 3편이 찾은 경로

위임 의사(303명)·대리인 비용(350개사)·알고리즘 관리(128명) — 신뢰는 자율성이 아니라 거버넌스에서 온다

4.1 **위임 의사의 형성 구조** 국내 학술지 3편, 2026

연구	표본·방법	핵심 경로와 계수
이종훈·안현철 **지능정보연구 32(1)**	303명 PLS-SEM	거버넌빌리티·책임성 명확성 → **AI 신뢰** → 태도·주관적 규범·지각된 행동통제 → 위임 의사. 행동의도에 가장 큰 영향은 **주관적 규범**(동료·상사의 기대). AI 대리성 지각은 지각된 위험을 키워 신뢰를 약화시키는 **부정 경로**. TPB 통합모형이 전통 TAM보다 설명력 우위
임원호 **Industry Promotion Research 11(3)**	350개사 설문	에이전틱 AI 도입 → 정보비대칭 감소 **β=.530**(p<.001), 감시효율성 향상 **β=.494**(p<.001) → 대리인 비용 감소(이중 매개). AI 거버넌스가 조절 효과 **β=.113**(p=.013)
강현정 **Information Systems Review 28(1)**	128명 PLS-SEM	인간-AI 협업(HAC)과 알고리즘 관리(AIgM)를 2차 요인으로 개념화. 둘 다 근접 AI 활용을 높이지만 **서로 다른 매개 경로**. 업무성과 예측력 Q^2=0.524

출처: 지능정보연구 32(1)·32(3), Industry Promotion Research 11(3), Information Systems Review 28(1) (2026)

① 신뢰는 '똑똑함'이 아니라 '설명 가능성'에서

두 국내 연구 모두 거버넌스(로그 추적·승인 체계·책임 소재)를 신뢰의 선행요인으로 확인했습니다. 성능 개선 캠페인보다 **책임 체계 문서화**가 위임률을 먼저 올립니다.

② 동료의 기대가 가장 강한 변수

주관적 규범의 계수가 가장 컸습니다. 위임은 개인 선호가 아니라 **조직 규범**의 함수입니다. 상사가 쓰는 팀에서 부하가 쓰지 않을 이유가 없고, 그 반대도 성립합니다.

③ 대리인 비용으로 번역하라

에이전트 도입의 재무 논리를 '인건비 절감'이 아니라 **정보비대칭·감시비용 감소**로 잡으면 투자 심의가 쉬워집니다. 단, 거버넌스가 없으면 조절효과가 사라집니다.

세 연구를 하나의 위임 방정식으로

국내 실증 3편은 서로 다른 표본을 다뤘지만 같은 구조를 확인합니다. **위임 의사 = (신뢰) × (규범) × (통제 가능성)**이며, 세 항 중 어느 하나가 0에 가까우면 전체가 0이 됩니다. 거버넌빌리티·책임 명확성은 신뢰를 만들고(β 경로), 동료·상사의 기대는 규범을 만들며, 정보비대칭·감시효율 개선은 통제 가능성을 만듭니다(β=.530·.494). 즉 'AI를 잘 쓰는 조직'과 'AI에게 일을 맡기는 조직'은 다르며, 후자는 **책임 구조를 먼저 정리한 조직**입니다.

경영진 체크리스트

① 에이전트의 결정에 **책임자를 이름으로** 지정했는가
② 로그로 **의사결정 과정을 재구성**할 수 있는가
③ 승인 체계가 **업무 위험도별로** 나뉘어 있는가
④ 위임 확대를 **성과 지표**로 선언했는가(비공식 실험 금지)
⑤ 실패 시 **롤백·재작업 비용**을 예산에 넣었는가

연구의 한계 세 연구 모두 **횡단 설문**이며 실제 위임 행동이 아니라 의사·인식을 측정했습니다. 계수는 경로의 존재를 보여줄 뿐 인과의 방향을 확정하지 않으므로, "거버넌스를 갖추면 위임이 늘어난다"는 해석은 종단·현장 실험이 필요한 가설로 다루어야 합니다. 그럼에도 세 표본(303명·350개사·128명)이 같은 방향을 가리킨다는 점은 실무 판단에 충분한 근거입니다.

V. 과학: 감독의 품질은 '통제량'이 아니라 '타이밍'이 결정한다

48명·192개 실시간 세션 실험: 계획 승인은 사고를 줄이지만 실행 중 개입을 줄인다

5.1 4가지 감독 전략의 실제 효과

arXiv:2604.04918

문제 행동 발생률 (낮을수록 좋음)



[그림 3] 계획 기반 전략 2종은 문제 행동 발생 오즈를 76% 감소(OR 0.24, 95% CI 0.14–0.42, p<.001). 출처: arXiv:2604.04918

그러나 문제 행동이 한 번 실행에 들어가면 **개입 성공률은 전략 간 차이가 없었고**(χ^2=3.17, p=.366), 전체 개입률은 7.4~26.2%에 그쳤습니다. 감독 전략은 '무엇을 시도하게 두는가'를 바꾸지만 '일어난 일을 사람이 잡아내는가'는 바꾸지 못했습니다.

왜 통제가 늘어도 못 잡는가

제어가 **너무 이르거나 너무 늦게** 돌아오는 타이밍 불일치. 제어가 제때 와도 사용자는 '안전한가'가 아니라 **'정답인가'**를 판단합니다. 사전 승인은 계획 밖 행동은 막지만, 계획에 없는 안전장치는 다루지 못하고 실행 중 감시를 오히려 억제합니다.

설계 원칙 3개

1. 권한·타이밍·주의를 **의사결정 분기점**에 맞춘다(모든 단계 확인이 아니라).
2. 승인 화면에 **'안전 판단'을 별도 항목**으로 분리한다(정답 판단과 혼동 방지).
3. 계획 승인과 실행 감시를 **같은 사람에게 겹쳐 놓지 않는다**.

5.2 네 전략의 작동 방식

Oversight strategy comparison

전략	작동 방식	관측된 강점·약점
행동 확인형	에이전트가 행동할 때마다 사용자가 확인	통제감은 높지만 피로가 누적되고 발생률은 83.3%로 높음
위험 게이트형	위험도가 높은 행동만 차단·확인	가장 높은 발생률(87.5%). 게이트 밖 행동이 사각지대로 남음
감독형 공동실행	계획을 함께 만들고 실행은 에이전트에 위임	발생률 56.2%로 최저. 실행 중 감시는 상대적으로 약함
구조화 강화형	계획에 안전 장치를 구조적으로 삽입	발생률 60.4%. 계획에 없는 안전장치는 여전히 비어 있음

출처: arXiv:2604.04918 Table 3·5. 48명·192세션, 4개 전략 무작위 배정. 개입률 차이는 통계적으로 유의하지 않음.

실무 번역 에이전트 자동화의 안전성은 '확인 창을 몇 개 띄우는가'로 측정되지 않습니다. **계획 단계에서 위험을 제거하고, 실행 단계에서는 되돌릴 수 있는 구간을 짧게 유지하며, 사후에는 로그로 재구성하는 3층 구조**가 실증적으로 지지됩니다.

VI. 교육: AI 튜터는 전문가를 따라잡았고, 표준시험은 따라잡지 못했다

1,179명·10개 영국 공립학교·10주 사전등록 RCT — 근접 효과는 확인, 광역 전이는 미확인

6.1 4개 실험군 · 3단계 학습 측정

Google × Eedi RCT, 2026-10

측정 단계	표준 보충학습	전문 인간 튜터	AI 튜터 (문제 맥락)	AI 튜터 (개인화 맥락)
즉시 오답 교정	43.4%	57.7%	**56.9%**	**56.5%**
근접 전이(다음 단원 첫 진단)	50.5%	59.1%	**58.4%**	**58.3%**
중간 퀴즈(8문항)	59.6%	61.8%	62.1%	**63.5%**
표준화 시험(Star Maths)	기준	-0.09 SD	-0.01 SD	+0.02 SD

출처: Google LeamLM, 「Measuring the impact of AI tutoring, from near transfer to standardised tests」(2026-10). 괄호는 95% 신뢰·베이즈 구간. 표준화 시험 차이는 모두 구간이 0을 포함.

품질: 사람 감독 하 AI는 '사람과 같은 정확도'

Gemini가 작성한 94,385건의 메시지 중 감독 튜터가 **94.7%를 수정 없이 승인**. 사후 감사에서 수학 오류는 1,000건당 3건 미만 — 전문 튜터와 같은 수준이었고, 어조 문제는 2건이 지적됐으나 학생에게 전달된 것은 없었습니다.

속도: 응답 지연이 '수업 시간'을 갈랐다

튜터 응답 대기 중앙값은 전문가 **24초**, AI(문제 맥락) 16초, AI(개인화) 19초. 대화 1회는 수업 시간 4~7분을 씁니다. 지연이 곧 커버리지 손실로 이어지는 구조입니다.

해석: '학습'의 정의가 문제다

좁은 지표에서 이긴 개입이 넓은 지표에서 사라지는 것은 흔한 일입니다. 이 RCT의 값은 'AI가 사람을 대체할 수 있는가'가 아니라 '**어느 지표를 목표로 설계할 것인가**'를 데이터로 분리했다는 데 있습니다.

6.2 실험 세부와 감독 비용

Design details

표본·기간

영국 공립 중등 10교, 8~10학년 1,179명. 주 1회 수학 수업 중 30~50분, 10주(시험 포함 12주). 4개 군 무작위 배정, 사전등록.

감독 비용

AI가 쓴 메시지 94,385건을 전문 튜터가 전건 검토. 94.7% 무수정 승인. 즉 **감독자의 시간이 병목**이며, 자동 승인 가능한 구간을 넓히는 것이 확장의 조건.

개인화 맥락의 효과

중간 퀴즈에서 개인화 맥락 군이 63.5%로 가장 높았고, 표준 대비 개선 확률 97.1%. 학습 이력 주입은 **비용이 거의 없고 효과가 가장 먼저 나타나는** 개입입니다.

근접 지표 (proximate)

플랫폼 안에서 본 즉시 교정·다음 단원 전이. 개입의 직접 효과를 가장 민감하게 포착하지만, 시험 점수와의 연결은 별도 문제입니다.

중간 지표 (intermediate)

교사가 직접 출제한 8문항 퀴즈. 4주간 배운 범위를 포괄해 근접과 광역 사이를 잇습니다.

광역 지표 (standardised)

외부 기관이 시행·채점한 Star Maths. 전 교육과정을 포괄하므로 **커버리지**의 영향을 강하게 받습니다.

VII. 교육: 마지막 10%는 커버리지에서 갈렸다 — 단원 132 vs 99 vs 77

표준시험 효과가 사라진 이유를 사후 분석이 추적: 대화가 깊을수록 범위가 얕아졌다

7.1 10주 동안 학생이 지나온 범위

Coverage analysis, exploratory

완료한 학습 단원 수(중앙값)



다룬 개념(construct) 수는 표준 72개, AI 57개, 인간 48개. 조건을 통제한 뒤에도 **단원 수 1 SD 증가 = 표준시험 +0.09 SD**(95% CI +0.03~+0.16). 즉 튜터링의 깊이는 넓이를 줄이고, 줄어든 넓이는 표준시험 점수로 나타납니다.

[그림 4] 출처: LeamLM RCT 부록 E·탐색적 분석(2026-10)

성별 이질성

탐색적 분석에서 표준시험 효과는 **여학생에서 더 우호적**이었고 남학생은 같은 광역 효과가 관측되지 않았습니다. 표본이 작아 확정은 이르지만, 개인화 맥락을 붙인 AI 튜터가 특정 집단에 비대칭적으로 작동할 가능성은 배포 설계에서 확인해야 할 항목입니다.

설계 함의

힌트(즉답) · 대화(4~7분) · 과제(자기주도)를 **주간 시간 배분 안에서** 섞어야 합니다. AI 튜터를 '모든 오답에 붙이는 버튼'으로 두면 근접 지표는 오르고 광역 지표는 정체합니다.

7.2 국내 교육 현장에 주는 세 가지

Korea application

정책: 연구 기반이 먼저 줄었다

2027년 예산안에서 국가AI교육연구센터 설립·운용 예산 30억 원이 전액 삭제됐습니다(전자신문 11-09). 이 RCT가 보여준 것은 효과 측정의 어려움이며, **측정 기관이 빠지는 순간** 현장 도입은 근접 지표(만족도·정답률)만으로 정당화됩니다.

학교: 커버리지를 시간표에 넣어라

전북교육청은 AI 리터러시 가이드라인 20개 조문 해설집을 전 학교에 배포했습니다. 가이드와 함께 **주당 최소 진도량**을 명시하면, 대화형 도구 도입이 진도를 잡아먹는 역설을 막을 수 있습니다.

기업 교육: 같은 논리가 적용된다

사내 AI 코칭도 '질문에 깊이 답하는 것'과 '과제를 완수하는 것'이 다릅니다. 교육 KPI를 **완료 과제 수·배포 건수**로 잡으면 같은 함정에 빠집니다.

핵심 AI 튜터의 근접 효과는 이미 사람 수준입니다. 다음 경쟁은 **깊이를 유지하면서 범위를 잃지 않는 수업 설계**이며, 이것은 모델 성능이 아니라 커리큘럼·시간표·평가의 문제입니다.

이 RCT가 남긴 세 가지 질문

질문	왜 중요한가
근접 효과는 누적되는가	40주(1학기) 이상 장기 추적이 없어, 10주 효과가 학기말 표준시험으로 합산되는지 알 수 없음
커버리지 손실은 설계로 줄일 수 있나	힌트·대화 비율을 조정한 군이 없어, "깊이와 범위"의 최적 배합이 아직 미지

읽는 법: "효과 없음"이 아니라 "다른 효과"

표준시험에서 차이가 없다는 결과는 AI 튜터가 무용하다는 뜻이 아닙니다. 같은 10주 동안 학생들은 좁고 깊게(오답 교정)와 넓고 얇게(진도)를 배운 것입니다. 평가가 무엇을 재는지에 따라 결론이 바뀌는 사례이며, **교육 구매 결정에서는 두 지표를 함께 계약 조건에 넣어야** 합니다.

Ⅷ. 교차 분석: 네 연구가 같은 병목을 가리킨다

기술·경영·과학·교육의 처방을 하나의 표로 정렬 — 무엇이 반복되고 무엇이 충돌하는가

8.1 연구 간 대조표 Cross-paper synthesis

쟁점	기술(멀티에이전트)	경영(위임·거버넌스)	과학(감독·협응)	교육(AI 튜터)
관측된 병목	조직화되지 않은 협업(해결률 6.25%)	위임 계약·책임 소재의 부재	타이밍 불일치, 실행 중 개입 실패	커버리지 손실(단원 -25%)
효과가 확인된 처방	역할 학습·상호작용 항·페일오버	증거 번들·변경 목록·한계 기술	계획 기반 승인(발생 -76%)	개인화 맥락·감독 하 생성
비용	토큰·지연 증가	토큰 +13%·시간 +38%	사전 승인이 실행 감시를 억제	수업 시간 4~7분/대화
측정되지 않은 것	장기 유지·조직 학습	실제 위임률 변화(횡단 설문)	장기 숙련·오탐 비용	광역 시험 전이(현재 0)
공통 결론	성능 상승은 이미 한계효용이 낮고, **검증·인수인계·커버리지 설계**의 한계효용이 높다			

반복되는 숫자: 20~40%

협업 점수 하락 20.2~46.6점, 위임 계약 비용 +13%/+38%, AI 튜터 커버리지 -25%. '검증과 구조'를 붙이는 비용은 대체로 10~40% 구간에 있고, 그 대가로 얻는 것은 **실패의 가시성**입니다.

충돌하는 지점

경영 연구는 '자율성 지각이 위험을 키운다'고 하고, 기술 연구는 '자율 조직화가 성능을 올린다'고 합니다. 모순이 아니라 **대상의 차이**입니다 — 조직화는 에이전트 간, 거버넌스는 사람-에이전트 간.

아직 비어 있는 칸

네 연구 모두 **3~10주** 관측입니다. 위임 계약이 1년 뒤 조직의 의사결정 품질을 올리는지, AI 튜터의 근접 효과가 누적되는지는 아직 데이터가 없습니다. 다음 분기의 실증 과제입니다.

산업별로 무엇을 먼저 바꾸는가

산업	가장 먼저 고칠 병목	첫 90일 과제
소프트웨어	인계 산출물 부재(변경 목록 7%)	PR 템플릿에 변경 파일·가정·한계 3종 고정
금융·보험	감시 효율과 책임 소재(β=.494)	승인 로그 표준화 + 위험도별 승인 단계 차등화
교육·에듀테크	커버리지 손실(단원 -25%)	주간 진도 목표와 대화 사용 비율 동시 설정
제조·물류	역할 페일오버 부재	에이전트 실패 시 대체 역할·수동 절차 문서화

독자가 오늘 할 수 있는 한 가지

지금 쓰는 AI 산출물 하나를 골라 **"다른 사람이 10분 안에 검증할 수 있는가"**를 물어보십시오. 아니라면, 이번 주에 필요한 것은 더 좋은 모델이 아니라 **변경 목록·가정·미해결 리스크를 적는 세 줄**입니다. 네 연구가 공통으로 지지하는 가장 값싼 개입입니다.

IX. 실무 프레임워크: 위임 등급제(Delegation Ladder) 5단계

연구 결과를 조직 운영 규칙으로 번역 — 무엇을 어떤 증거와 함께 넘길 것인가

9.1 위임 등급과 요구 증거

Operational design

L1 초안 생성 — 사람이 전부 작성
에이전트는 초안만. 산출물에 출처·가정 목록 필수. 검토자는 사실 확인만.

L2 제안 + 검토 체크리스트
변경 파일·미해결 리스크·알려진 한계 3종을 산출물에 강제. 위임 계약 연구의 최소 번들.

L3 실행 + 사전 계획 승인
계획을 먼저 승인받고 실행. 문제 행동 -76%의 효과를 얻는 대신, 실행 중 감시자를 따로 둔다.

L4 자율 실행 + 사후 감사
로그·샘플 감사. 감사 표본률과 롤백 절차를 사전에 합의. 거버넌스 성숙 기업의 표준.

L5 자율 실행 + 자율 복구
역할 페일오버·자가치유. 고빈도·저위험 업무에만. 판단이 걸린 업무는 L4에서 멈춘다.

등급을 올리는 조건

① 최근 30건에서 숨은 인수 테스트(또는 동등 검증) 통과율 ≥ 95%
② 검토자 판단 시간이 3개월간 감소 추세
③ 사고 1건당 복구 시간(MTTR)이 목표 내
④ 로그로 재구성 가능한 비율 100%

학교·기업의 교육 설계

AI 튜터 도입 시 **주간 커버리지 목표**를 함께 설정합니다(예: 단원 12개/주, 힌트·대화 비율 7:3). 근접 지표와 광역 지표를 같은 대시보드에 두고, 광역 지표가 2주 연속 정체하면 대화 비율을 낮춥니다.

측정 지표 4종 + 재심사 원칙

검토가능성(리뷰어 모호성), 인수인계 완결률(증거 번들 충족), 개입률(문제 행동 중 사람이 멈춘 비율), 커버리지(단위 시간당 처리 범위). 등급은 **업무 단위**로 부여하고 **분기마다 한 단계씩 재심사**합니다 — 같은 계약에서 통과율이 오르는 것이 상향의 유일한 근거입니다.

9.2 업무별 배정 예시

Illustrative assignment

업무	등급	요구 증거	승인·감사
공개 문서·보고서 작성	L2	출처 목록·가정·미확인 항목	담당자 검토 후 발행
고객 응대 초안	L2	정책 인용·예외 처리 근거	팀장 승인
내부 데이터 집계·리포트	L4	쿼리 로그·검증 샘플	주간 샘플 감사 10%
코드 변경·배포	L3	변경 목록·테스트·롤백 계획	PR 리뷰 1인 + 자동 revert

콘데리 제안 예시. 등급 기준은 조직의 위험·규제 수준에 맞춰 조정해야 함.

X. 결론: 다섯 문장과 세 개의 관측 지표

오늘의 학술 신호를 창업가·경영진·교육자가 바로 쓸 수 있는 형태로 압축

다섯 문장

1. 멀티에이전트의 성능 상한은 모델 능력이 아니라 조직 설계가 정한다 — MAC-Bench 최고 6.25%.
2. 위임 계약은 정답률을 올리지 않는다. 검토가능성을 +0.83 올리고, 그 대가로 토큰 13%·시간 38%를 쓴다.
3. 사람의 위임 의사는 자율성 인식이 아니라 거버넌스·책임 명확성에서 나온다(303명·350개사 실증).
4. 감독은 게이트 수가 아니라 타이밍이다. 사전 승인은 발생을 76% 줄이지만 실행 중 개입은 7.4~26.2%에 머문다.
5. AI 튜터는 근접 효과에서 전문가와 동등하고, 광역 효과는 커버리지 설계에 달려 있다 — 단원 132 vs 99 vs 77.

결론 지금 가장 값싼 성능 개선은 모델 교체가 아니라 **인수인계 산출물의 표준화**다. 그것이 없으면 어떤 벤치마크 점수도 조직의 성과로 번역되지 않는다.

관측 지표 ① 리뷰어 모호성

검토자가 "이게 맞나?"라고 되묻는 비율. 위임 계약 연구에서 유의하게 감소한 유일한 주관 지표.

관측 지표 ② 증거 충족률

변경 목록·한계·리스크 3종을 갖춘 산출물 비율. 계약 없이는 7%·2%에 머물렀습니다.

관측 지표 ③ 커버리지

단위 시간당 처리한 범위. 교육에서는 단원 수, 업무에서는 완료 케이스 수. 깊이의 대가를 재는 유일한 눈금.

다음 호 예고

에이전트 조직의 장기 학습(채봉석·박은혜 3계층 아키텍처) 후속 실증과, 위임 계약의 6개월 효과를 추적합니다.

참고문헌 / References

[1] S. Wu, "Measuring AI Agents as Teammates with Collaborative Intelligence," arXiv:2611.02633, 2026-11-02.
[2] K. Jha, "s3: Simple Scaling of Swarms via Cooperative Multi-agent RL," arXiv:2611.08615, 2026-11-05.
[3] Ahn et al., "MA-Mesh: Resilient Multi-Agent Collaboration via Role-based Self-Healing," arXiv:2611.06389, 2026-11-04.
[4] Z. Chen, "MAC-Bench: Benchmarking Multi-Agent Collaboration via Interconnected Tasks," arXiv:2611.01728, 2026-11-07.
[5] "Learning Multi-Agent Orchestration: Higher-Order Interactions in Policy Improvement (IAPO)," arXiv:2611.07495, 2026-11-05.
[6] S. Agashe, Y. Fan, A. Reyna, X. E. Wang, "Zero-Shot Coordination among LLM Agents," arXiv:2611.06409, 2026-11-04.
[7] "Software Delegation Contracts: Measuring Reviewability in AI Coding-Agent Work," arXiv:2606.17099, 2026-06-14.
[8] C. Chen et al., "Comparing Human Oversight Strategies for Computer-Use Agents," arXiv:2604.04918, 2026.
[9] Google LearnLM, "Measuring the impact of AI tutoring, from near transfer to standardised tests," 2026-10.
[10] 이종훈·안현철, 「Agentic AI에 대한 조직 구성원의 위임 의사 형성 메커니즘」, 지능정보연구 32(1), 2026.
[11] 임원호, 「에이전틱 AI 도입이 조직 대리인 비용 감소에 미치는 영향」, Industry Promotion Research 11(3), 2026.
[12] 박호연·김경재, 「단일·멀티 에이전트 금융투자보고서 생성 비교 평가」, 지능정보연구 32(3), 2026-09-30.
[13] 강현정, 「The Impact of Algorithmic Management and Human-AI Collaboration」, Information Systems Review 28(1), 2026.
[14] 채봉석·박은혜, 「An Agentic AI System to Improve Organizational Learning and Memory」, 한국경영정보학회 춘계학술대회, 2026-06.