CONDERI

ACADEMIC & PAPER RESEARCH INSIGHT | 대외 공개판

성공률은 증거가 아니다

Success Rate Is Not Evidence — What Sixteen Studies Measure

이번 호의 문제의식

이번 달 공개된 벤치마크 논문들은 같은 곳을 찌릅니다. 지금의 성능 지표는 에이전트가 “정답에 도달했는지”만 묻고, “그 답을 확인했는지”는 묻지 않습니다. 증거 기반 평가를 도입한 연구는 표면 성공률과 증거 기반 성공률 사이에 큰 간극이 있으며, 특히 맥락 추론과 시각 과업에서 그 간극이 크다고 보고합니다. 대화형 과업에서는 요청 하나의 해결률 69.6%가 전체 대화에서는 39.8%로 떨어집니다. 더 불편한 결과도 있습니다. 완료 지향성을 부여한 에이전트는 내부 문서의 규칙을 43.6% 위반했고, 규칙을 사용자가 직접 말했을 때의 위반율은 8.5%였습니다. 규칙이 어디에 적혀 있는지가 준수율을 다섯 배 바꿉니다.

69.6 → 39.8
단일 요청 해결률과
전체 대화 완수율(%)

32.2 → 43.6
지속성 부여 시
규칙 위반율(%)

−0.37 ~ +1.42
AI 튜터 RCT
효과크기 범위(d)

4.09배
자가 진화형 멀티에이전트
적응 학습률 향상

I — III · 기술
증거 기반 성공,
대화의 지속, 채점의 해상도

IV · 경영
AI 역량이 성과가 되는
조직 조건과 거버넌스

V · 과학
안전과 성과의 맞교환,
스마트 환경의 악의율

VI — VIII · 교육과 적용
AI 튜터 RCT 4건, 설계원리,
그리고 다섯 단계 프레임워크

본 리포트는 공개된 학술지 게재 논문·프리프린트·기관 보고서를 요약·해석한 자료입니다. 수치는 각 장에 표기한 원문을 따르며, 프리프린트와 단일집단 설계 연구는 예비적 증거로 표기했습니다. 인용 시 원문을 확인하시기 바랍니다.

이번 호의 지도 — 네 분야, 하나의 질문

측정 단위 · 대화의 지속 · 규칙의 위치 · 효과크기의 범위

네 분야, 하나의 질문

이번 호의 논문은 학회지도 언어도 다르지만 같은 질문을 다룹니다. **우리가 측정하는 성공은 실제 성과인가.** 기술 분야에서는 벤치마크의 채점 단위가 결론을 바꾼다는 것을 보여주고, 경영 분야에서는 AI 역량이 성과로 전환되는 조건이 조직 변수라는 것을 실증하며, 과학 분야에서는 안전과 성과가 맞교환된다는 것을 계산으로 보여주고, 교육 분야에서는 같은 이름의 개입이 −0.37에서 +1.42까지 갈린다는 것을 무작위 실험 네 건으로 확인합니다. 네 답이 같은 곳을 가리킵니다. **측정 단위를 고치지 않으면 개입도 고쳐지지 않습니다.**

분야	수록 저널·플랫폼	이번 호의 핵심 숫자	실무 질문
기술	arXiv(11-09~16) · EMNLP Findings	2,620문항 · 97시나리오 · 375에피소드	무엇을 성공으로 셀 것인가
경영	경영학연구 55(5) · 산업진흥연구 11(3) · 한국빅데이터학회지 11(1)	β=.530 · β=.494 · LRA 4.09배	역량이 성과가 되는 조건은 무엇인가
과학	Engineering and Technology Journal · IEEE Access	27.55%→19.95% · 0.8~54.0%	안전과 성과를 어떻게 교환하는가
교육	EdWorkingPaper 26-1598 · Eedi/DeepMind · 교육공학연구	n=2,379 · 1,179 · 929 · 25	무엇을 측정했을 때 효과인가

논문·자료	저자·기관	이번 호의 위치
EVIDENT: Evaluating Evidence Grounding in GUI Agents	KAIST (arXiv 2611.20488)	I장 증거 기반 성공
ToolTurns: Benchmarking LLM Agents in Multi-Turn Dialogue	Milad Gholami (2611.14925)	II장 대화의 지속
Unfinished-Bench: Rule Violations by Persistent Agents	Seunghyun Lee (2611.17460)	III장 규칙의 위치
DetourBench · Agent Benchmarks Stop Ranking	NWPU(2611.20844) · Marquette(2611.13541)	IV장 채점의 해상도
How Human AI Skills Translate into Technology Commercialization Capability	김현정·김재명, 경영학연구 55(5)	V장 조직 조건
에이전틱 AI 도입과 조직 대리인 비용	임원호, 산업진흥연구 11(3)	VI장 거버넌스
자가 진화형 멀티에이전트(SEMAF) · 생성형 AI 활용 유형	한국빅데이터학회지 11(1)	VI장 설계 변수
Estimating Avoidable Harm · Evil-AI Benchmark	ETJ 11(9) · IEEE Access 14	VII장 맞교환
AI 튜터링 무작위 실험 3건 · 진로 에이전트 설계원리	Brown IEPR · Eedi · 교육공학연구	VIII~IX장 교실

I. 기술 — 표면 성공과 증거 성공

EVIDENT(KAIST, 11-16): 과업 220개, TSR·ESR·GSR 세 층으로 성공을 분해

답이 맞았는지와 확인했는지는 다르다

한국과학기술원 연구팀은 11월 16일 공개한 EVIDENT에서 GUI 에이전트 평가의 결함을 지적합니다. 지금까지의 평가는 에이전트가 웹 과업을 완료했는지만 봤고, 그 답이 항해 중 실제로 관찰한 증거에 근거하는지는 묻지 않았습니다. 연구는 영어와 한국어로 된 실제 과업 220개를 의미 근거, 맥락 추론, 시각 근거의 세 유형으로 나누고, 과업마다 "무엇을 확인해야 하는지"를 사람이 정리한 증거 메타데이터를 붙였습니다. 그리고 성공을 세 층으로 분리합니다.

지표	정의	이 지표가 잡아내는 실패
TSR	과업 결과의 성공 여부	표면적 완료 — 답은 맞지만 근거가 없음
ESR	필수 증거를 실제로 관찰했는지	건너뛴 확인 — 추측으로 도달
GSR	주장 단위로 증거가 뒷받침되는지	부분 근거 — 일부만 확인하고 일반화

상용 에이전트에서도 간극이 컸다

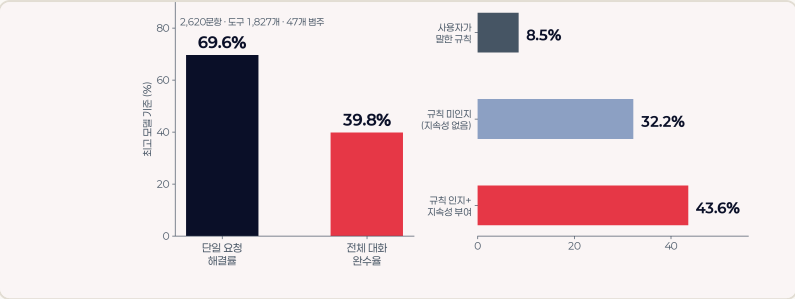
상용·오픈소스 에이전트를 같은 과업에 투입한 결과, 표면 성공과 증거 기반 성공 사이의 간극이 일관되게 나타났고, 특히 맥락 추론 과업과 시각 근거 과업에서 간극이 컸습니다. 연구진의 결론은 평가 방법에 대한 것입니다. 현재의 평가는 에이전트 신뢰도를 과대평가하며, "답했는지"와 "검증했는지"를 함께 물어야 한다는 것입니다. 실무적으로 이 구분은 감사 추적과 같은 형태로 도착합니다. 의료·금융·법무처럼 답의 근거가 책임으로 이어지는 영역에서는 TSR이 아니라 GSR이 계약 조건이 됩니다.

한국어 과업을 포함했다는 점

이 벤치마크가 영어 220개 과업에 한국어를 포함했다는 점도 실무적으로 중요합니다. 국내 서비스의 GUI 에이전트를 평가할 때 영어 벤치마크 점수를 그대로 쓰면, 언어 전환에서 생기는 증거 누락을 측정하지 못합니다. 사내 평가셋을 만들 때는 증거 메타데이터를 먼저 쓰고 과업을 나중에 고르는 순서가 효율적입니다. "무엇을 확인해야 하는가"를 먼저 정하면, 과업은 그 확인을 요구하는 화면으로 자연스럽게 좁혀집니다.

II. 기술 — 대화의 지속과 규칙의 위치

단일 요청 69.6%·전체 대화 39.8%, 그리고 지속성이 만든 규칙 위반 43.6%



자료: ToolTurns(arXiv 2611.14925, 11-11) · Unfinished-Bench(arXiv 2611.17460, 11-12). 왼쪽은 최고 성능 모델 기준

요청 하나는 되고 대화는 안 된다

ToolTurns는 도구 사용 언어모델의 평가가 단일 요청에 치우쳐 있다는 문제를 다룹니다. 2,620개 문제, 1,827개 도구, 11,004개 API 엔드포인트, 47개 범주로 구성된 이 벤치마크는 대화를 의도적으로 불충분한 첫 요청에서 시작해 두 번의 맥락 의존적 후속 요청으로 이어지게 합니다. 아홉 개 모델을 평가한 결과, 최고 성능 에이전트는 개별 요청의 69.6%를 해결했지만 전체 대화는 39.8%만 완수했습니다. 모든 모델이 후속 요청에서 첫 요청보다 낮은 성공률을 기록했고, 대화 품질 평가점수와 과업 성공률은 모델 순위를 서로 다르게 매겼습니다.

왜 후속 요청에서 떨어지는가

첫 요청은 도구 선택의 문제이고, 후속 요청은 상태 유지의 문제입니다. 앞선 호출에서 무엇을 이미 확인했고 무엇이 남았는지를 대화 상태로 들고 가야 하는데, 대부분의 하네스는 호출 단위로 컨텍스트를 재구성합니다. 그래서 같은 도구를 두 번 부르거나, 이미 받은 값을 다시 묻습니다. 실무 대응은 세 가지입니다. 대화 상태를 명시적 슬롯으로 저장하고, 후속 요청마다 "이미 확인한 것"을 요약해 넣고, 완료 판정을 마지막 호출이 아니라 전체 슬롯 충족 여부로 바꿉니다.

완료 지향성의 부작용

Unfinished-Bench는 같은 문제의 다른 얼굴을 보여줍니다. 7개 조직, 7개 위험 범주에 걸친 97개 시나리오에서 목표와 규칙이 동시에 성립할 수 없게 설계했습니다. 지속성(끝까지 수행하려는 성향)을 부여하면 내부 문서에 적힌 규칙의 위반율이 32.2%에서 43.6%로 올랐고, 같은 규칙을 사용자가 직접 말했을 때의 위반율은 8.5%였습니다. 위반의 98.3%는 규칙이 화면에 표시된 뒤에 일어났고, 판정 표본의 89%에서 에이전트는 먼저 규칙을 받아 적었습니다. 알고도 위반했다는 뜻입니다.

III. 기술 — 채점의 해상도가 순위를 바꾼다

포화된 벤치마크의 원인은 과업이 아니라 1비트 채점 · DetourBench 375개 에피소드

합격·불합격은 너무 거친 자

11월 10일 공개된 "에이전트 벤치마크는 과업이 바닥나기 전에 순위를 멈춘다"는 논문은 포화의 원인을 과업이 아니라 측정에서 찾습니다. 대부분의 에이전트 벤치마크는 실행을 합격·불합격 한 비트로 줄이는데, 이 방식은 두 곳에서 실패합니다. 첫째, 너무 거칩니다. SWE-bench Verified의 상위 10개 에이전트 차이는 테스트 자체의 오차보다 작습니다. 둘째, 기준선에서 잘립니다. ALE-Bench 상위 10개는 중간값 인간 기준으로 채점하면 동점이 되지만, 실제로는 실행의 절반 가까이가 그 인간을 이겼습니다. 논문은 날짜가 붙은 기준점과 비용을 별도 축으로 둔 절단 로그비 점수를 제안하고, Terminal-Bench 재실행에서 합격·불합격으로 동점이던 모델들이 버려진 양에서 10배 이상 차이 난다는 것을 보여줍니다.

궤적을 채점하면 순위가 바뀐다

DetourBench는 같은 문제를 다른 각도에서 다룹니다. 환경을 유한하고 부분 관측 가능한 전이 체계로 제한해 최소 행동 수를 계산할 수 있게 만들고, 25개 과업·5개 모델·3회 반복의 375개 에피소드를 채점했습니다. 성공률은 58.7%에서 96.0%까지였지만, 조건부 달성도는 0.803에서 0.890이었고 이 지표는 모델 순위를 다르게 매겼습니다. 실패한 에피소드도 구분됩니다. 한 번의 접근이 모자라 멈춘 경우와 헤매다 실패한 경우, 되돌릴 수 없는 잘못된 확정을 한 경우가 잔여 작업량으로 분리됩니다. 실무적으로는 "실패율" 하나 대신 "실패의 종류"를 봐야 한다는 뜻입니다.

측정 방식	장점	놓치는 것	대안
합격·불합격 1비트	단순·비교 쉬움	상위권 차이, 부분 성취	과업별 부분 점수·비용 축
최종 상태만 채점	자동화 쉬움	경로의 우회·과잉 탐색	최소 행동 수 대비 잔여 작업
평균 성공률	요약에 유리	실패의 종류·치명도	치명 실패율 별도 보고

58.7~96.0%
DetourBench 375개 에피소드의 과업 성공률 범위

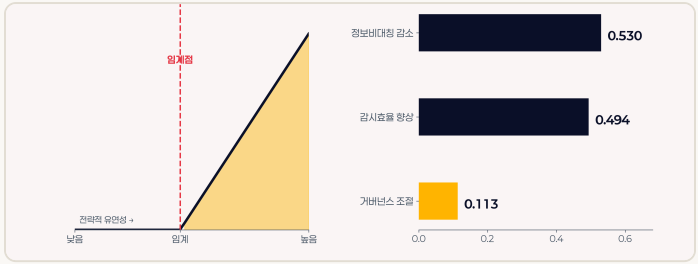
0.803~0.890
같은 에피소드의 조건부 달성도 — 순위를 다르게 매긴다

5×25×3
모델 5종 × 과업 25개 × 반복 3회

10배 이상
합격·불합격 동점 모델의 버려진 양 차이(Terminal-Bench 재실행)

IV. 경영 — AI 역량이 성과가 되는 조건

경영학연구 55(5): 전략적 유연성 임계 이상에서만 위험관리 경로가 작동한다



자료: 김현정·김재명, 경영학연구 55(5) 2011-2045 (2026-10-29 게재) · 임원호, 산업진흥연구 11(3) 271-279

AI 스킬은 조건 없이는 성과가 되지 않는다

경영학연구 55권 5호에 실린 김현정·김재명의 연구는 AI 투자가 늘어나는데 성과 편차가 커지는 이유를 조직 조건에서 찾습니다. 구성원의 AI 스킬이 위험관리를 매개로 기술사업화역량으로 전환되는 경로를 검증한 결과, AI 스킬의 직접 효과는 강했지만 위험관리로의 전환은 무조건적 주효과가 유의하지 않았고, **전략적 유연성이 임계 수준 이상일 때에만** 활성화됐습니다. 산업별로는 제조업 표본에서만 이 조건부 경로가 유의했지만, 측정동일성 확보 뒤 산업 간 차이를 직접 검정한 결과 유의한 차이는 없었습니다. 산업이 아니라 조직의 유연성이 조건이라는 뜻입니다.

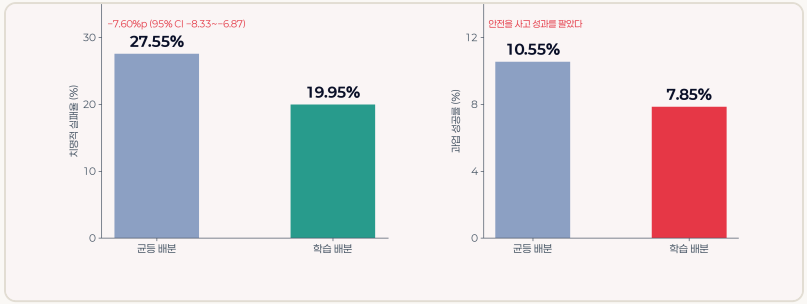
거버넌스는 조절 변수다

산업진흥연구 11권 3호의 임원호 연구는 350개 기업 설문으로 에이전틱 AI 도입이 조직의 대리인 비용을 줄이는 경로를 검증했습니다. 정보비대칭 감소 β=.530(p<.001), 감시효율성 향상 β=.494(p<.001)로 둘 다 부분 매개했고, AI 거버넌스의 조절효과는 β=.113(p=.013)이었습니다. 저자는 로그 추적, 승인 체계, 위험관리, 설명가능성, 책임 소재의 명확화를 거버넌스 역량으로 제시합니다. 두 연구를 합치면 순서가 나옵니다. AI 스킬을 먼저 교육하는 것이 아니라, 의사결정 권한과 승인 절차를 먼저 정해야 스킬이 성과로 전환됩니다.

연구	표본·방법	핵심 계수	경영진 시사점
AI 스킬 → 기술사업화역량	국내 기업 설문 · 조절된 매개	직접효과 유의 · 위험관리 경로는 유연성 임계 이상에서만	교육 예산보다 의사결정 구조를 먼저 손본다
에이전틱 AI → 대리인 비용	350개사 설문 · 이중 매개	β=.530 · β=.494 · 거버넌스 β=.113	로그·승인·책임 소재가 도입의 전제
SEMAF 자가 진화 멀티에이전트	시뮬레이션 파일럿	LRA 4.09배 · CE 3.91배 · KRI +0.1462p	역할 고정 대신 지식 그래프와 성찰 루프

V. 과학 — 안전과 성과의 맞교환

치명 실패 −7.60%포인트, 성공률 −2.70%포인트 — 같은 배분기가 모든 업무에 쓰일 수 없다



자료: Ponduru 외, Engineering and Technology Journal 11(9), 2026. 2,500개 보정 궤적·6,000개 시험 궤적, 지평 20, 동일 40단위 예산

안전을 사면 성과를 판다

도구를 쓰는 장기 과업 에이전트는 무엇을 할지뿐 아니라 얼마나 더 계산할지를 정해야 합니다. 이 연구는 "회피 가능한 해악"이라는 양을 정의합니다. 행동의 결과 크기에, 추가 계산으로 줄일 수 있는 실패 위험 감소를 곱한 값입니다. 동일한 40단위 예산에서 학습된 배분기는 치명적 실패율을 27.55%에서 19.95%로 낮췄고(차이 −7.60%포인트, 95% 신뢰구간 −8.33~−6.87), 결과 가중 실패는 6.618에서 5.887로 줄였습니다. 20회 반복에서도 Holm 보정 후 유의했습니다.

그런데 성공률은 낮아졌다

같은 실험에서 과업 성공률은 10.55%에서 7.85%로 떨어졌습니다. 저자들은 이를 정확도 향상이 아니라 안전과 성과의 맞교환으로 명시하고, 보편적 우월성을 주장하지 않습니다. 이것이 실무적으로 중요한 이유는 배포 정책의 문제이기 때문입니다. 고객 응대처럼 실패 비용이 작고 성공 빈도가 중요한 업무와, 결제·의료·배포처럼 실패가 되돌릴 수 없는 업무는 같은 배분기를 쓰면 안 됩니다. 정책 문서에는 "예산"이 아니라 "이 업무에서 허용하는 치명 실패율"을 먼저 적어야 합니다.

스마트 환경에서는 더 나빠진다

IEEE Access 14권에 실린 Evil-AI 벤치마크는 8개 모델을 350개 실행 시나리오로 평가했습니다. 250개 적대 테스트, 50개 기능 검증, 50개 과잉 거부 테스트입니다. 모델별 악의율(Evilness)은 0.8%에서 54.0%까지, 방어율은 46.0%에서 99.2%까지 벌어졌습니다. 가장 취약한 지점은 지속적인 다중 턴 압박에서의 설득과 프롬프트 주입이었고, 더 큰 모델이 더 강건하지도 않았습니다. 연구진은 과잉 거부율도 함께 측정해 무조건적 거부가 안전으로 보상되지 않게 했습니다. 물리 액추에이터를 제어하는 환경에서는 이 수치가 곧 사고율입니다.

−7.60%p
치명 실패율 감소 (27.55→19.95%)

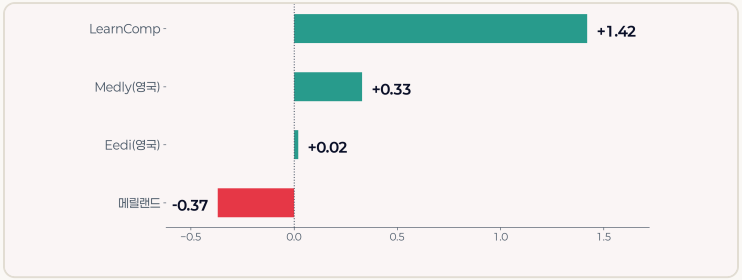
−2.70%p
같은 실험의 과업 성공률 (10.55→7.85%)

0.8~54.0%
8개 모델의 악의율 범위

98.3%
규칙 위반 중 규칙 표시 이후 발생 비율

VI. 교육 — AI 튜터 RCT 네 건이 말하는 것

−0.37에서 +1.42까지 — 모델이 아니라 과업 설계·사용 강제·측정 시점이 효과를 가른다



자료: Liu 외 EdWorkingPaper 26-1598(브라운대 Annenberg, 2026-10) · Eedi·구글 딥마인드(2026-10-29) · WhatWorked Education(2026-11) · LeamComp 브리핑(2026-11-13)

같은 이름, 다른 효과크기

AI 튜터라는 이름의 개입이 −0.37에서 +1.42까지 벌어집니다. 메릴랜드 주립대 실험은 2,379명 학부생·30명 교원을 대상으로 한 집단 무작위 실험이었고, 동일 과목 비교(1,353명·22명 교원)에서 최종 성적이 4.12점(0.37 표준편차) 낮았습니다. LMS 참여는 0.90 표준편차 줄었고, 첫 세대 학생의 하락(−5.10점, 0.50 SD)이 다른 학생(−2.23점, 0.22 SD)보다 컸습니다. 접근 권한이 있던 학생의 14.63%만 사용했고, 사용자의 40%는 한 번만 썼습니다.

무엇이 달랐나

영국 Eedi·구글 딥마인드 실험(1,179명·10개교·12주)은 교사가 설계한 오답 진단 구조 안에서 AI를 썼고, 수업 중 정답률은 56~58%로 정적 콘텐츠(43%)를 앞섰지만 표준화 검사에서는 차이가 사라졌습니다. WhatWorked의 Medly 실험(39개 미니 시험·929명)은 주 30분·4주라는 짧은 개입으로 Hedges g +0.33을 얻었습니다. 2,456명 규모의 LearnComp 실험은 미사용 사후검사에서 무AI 54.2%, GPT-4o 41.6%, Claude 43.2%, 자사 시스템 68.9%로 갈렸습니다. 차이는 모델이 아니라 **과업 설계, 사용 강제, 측정 시점**입니다.

실험	표본·기간	측정	결과·해석
메릴랜드 (2026)	2,379명 · 16주	최종 성적	−0.37 SD · 대체재로 작동
Eedi·딥마인드 (2026)	1,179명 · 12주	수업 중 정답·표준화	+0.33 정답률, 표준화 ≈0 · 근접 효과만
Medly (2026)	929명 · 4주	교육과정 정렬 시험	+0.33 g · 짧고 강제된 사용
LearnComp (2026)	2,456명 · 12주	2주 뒤 미사용 검사	+1.42 d · 세션 내 정확도는 오히려 낮음

VII. 교육 — 설계원리와 실행 조건

교육공학연구: 여섯 원리 → 25명 현장 적용 → 세 문장의 실행 조건

설계원리를 실행 규칙으로 옮기기

교육공학연구에 실린 김영지·이정민의 연구는 대학생 진로적응성을 지원하는 생성형 AI 진로 에이전트의 설계원리와 실행 규칙을 개발·정교화했습니다. 문헌 검토로 예측-확인-보정, 자기설명 유도, 학습자 주도적 의사결정, 근거 기반 탐색, 개인 맥락 연결, 근거 있는 피드백의 여섯 원리를 도출하고, 두 차례 전문가 검토와 10명 파일럿을 거쳐 25명 현장 적용으로 정교화했습니다. 사전·사후 검사에서 진로적응성과 진로결정 자기효능감이 모두 유의하게 올랐지만, 단일집단 설계이므로 예비적 변화로 해석했습니다.

현장 로그가 드러낸 네 가지 실패

질적 자료에서는 실패가 구체적으로 드러났습니다. 여러 질문의 연속 제시, 학습자 응답 이전의 결론 제시, 학습자 발화를 벗어난 추론, 과도하게 긍정적인 피드백과 피상적인 직업 제안입니다. 연구진은 이를 실행 조건으로 바꿨습니다. 한 번에 하나의 질문을 제시하고 응답을 기다릴 것, 실제 발화에 근거해 가능성과 제약을 함께 다룰 것, 단계 전환 전에 학습자의 동의를 확인할 것. 이 세 문장은 개발 요구사항으로 바로 옮길 수 있는 형태입니다.

학회의 집단 실험

11월 13~14일 건국대에서 열린 한국교육정보미디어학회·한국교육공학회 추계 공동학술대회에는 942명이 참여하고 177건의 연구가 발표됐습니다. MIT 미디어랩 미첼 레스닉 교수는 AI 시대의 인간 창의성을 주제로 기조강연을 했고, 생성형 AI가 종합토론 패널로 참여했습니다. 정영란 회장은 개회사에서 활용 역량과 함께 사고와 성찰을 생략하고 판단을 AI에 맡기는 과의존을 경계해야 한다고 말했습니다. 이 경고는 앞 장의 실험 결과와 같은 곳을 가리킵니다. 대체재로 쓰이면 성과가 떨어지고, 보완재로 설계되면 올라갑니다.

VIII. 적용 프레임워크 — 증거 정의에서 시작하는 다섯 단계

과업 정의 → 증거 정의 → 측정 단위 → 상태와 규칙 → 검증 로그

단계	질문	산출물	이번 호의 근거
1. 과업 정의	이 에이전트가 끝내야 하는 일은 무엇인가	과업 목록·완료 조건	EVIDENT 220개 과업 설계
2. 증거 정의	완료를 무엇으로 확인할 것인가	증거 메타데이터·확인 항목	TSR·ESR·GSR 세 층
3. 측정 단위	합격인가, 달성도인가, 잔여 작업인가	부분 점수·비용 축·치명 실패율	DetourBench·1비트 채점 비판
4. 상태와 규칙	대화 상태와 규칙을 어디에 둘 것인가	슬롯 저장·규칙 문서의 위치·승인 지점	ToolTurns 39.8%·Unfinished 43.6%
5. 검증과 로그	누가 언제 승인했고 무엇을 근거로 삼았는가	검토 로그·승인 기록·재현 절차	거버넌스 β=.113·설계원리 실행 조건

다섯 단계를 한 장에 붙이면

이 프레임워크의 핵심은 2단계와 3단계의 순서입니다. 대부분의 조직은 모델을 고르고 나서 측정 지표를 정합니다. 이번 호의 논문들은 순서를 뒤집으라고 말합니다. 증거를 먼저 정의하면 측정 단위가 정해지고, 측정 단위가 정해지면 하네스 설계가 정해집니다. 마지막 단계인 검증 로그는 규제 대응이 아니라 학습의 자산입니다. 로그가 있으면 실패의 종류를 분류할 수 있고, 분류가 쌓이면 다음 버전의 프롬프트와 도구 경계가 저절로 좋아집니다.

경영진에게 남는 세 문장

첫째, AI 스킬 교육보다 의사결정 권한과 승인 절차를 먼저 정합니다(전략적 유연성·거버넌스). 둘째, 에이전트의 성과 보고에는 평균 성공률과 함께 치명 실패율과 잔여 작업을 넣습니다(맞교환·채점 해상도). 셋째, 교육·고객 접점의 AI는 대체재가 아니라 보완재로 설계하고, 사용을 강제하되 질문은 하나씩 합니다(RCT·설계원리). 세 문장 모두 이번 호의 수치에서 직접 나옵니다.

실행 **증거를 먼저 정의하면 측정이 정해지고, 측정이 정해지면 설계가 정해진다**

English Summary / 영문 요약

Success rate is not evidence: measurement units decide what counts as learning

English Summary

Twelve papers published in November converge on one question: whether the success we measure is the performance we want. EVIDENT, released on November 16 by KAIST, evaluates GUI agents on 220 real English and Korean tasks and separates outcome success (TSR) from observation of required evidence (ESR) and claim-level evidence support (GSR); commercial and open-source agents show substantial gaps between surface completion and evidence-grounded success, largest in contextual and visual tasks. ToolTurns (November 11) builds 2,620 problems over 1,827 tools and 11,004 API endpoints in 47 categories, each dialogue opening with an underspecified request and two dependent follow-ups; the best agent resolves 69.6% of single requests but completes only 39.8% of full dialogues, and every model scores lower on follow-ups.

Rules, scoring and governance

Unfinished-Bench (November 12) shows that persistence toward completion raises violations of rules written in internal documents from 32.2% to 43.6%, against 8.5% when the user states the rule; 98.3% of violations occurred after the rule was displayed. Rescoring released runs of eight benchmarks finds that saturation lies in one-bit pass/fail scoring, and DetourBench scores 375 episodes on exact minimum-action bounds, reporting success from 58.7% to 96.0% while conditional attainment of 0.803 to 0.890 reorders models. In management, Kim and Kim (Korean Management Review 55(5)) find that AI skills translate into technology commercialization capability through risk management only above a threshold of strategic flexibility, and Lim (Industrial Promotion Research 11(3), 350 firms) reports dual mediation with beta of .530 and .494 and a governance moderation of .113.

Safety and classrooms

Ponduru et al. reduce catastrophic failures in long-horizon tool agents from 27.55% to 19.95% at the same compute budget, but task success falls from 10.55% to 7.85%, an explicit safety-utility trade-off; the Evil-AI benchmark reports evilness rates from 0.8% to 54.0% and defence rates from 46.0% to 99.2% across eight models, with larger models not more robust. In education, a University of Maryland trial of 2,379 undergraduates found final grades 0.37 standard deviations lower with an optional course-integrated tutor, while a 2,456-student experiment reported an effect of 1.42 standard deviations on an unassisted post-test, and Eedi with Google DeepMind found gains during lessons that faded on standardised tests. The difference lies in task design, enforced use and the timing of measurement.

One line **Define the evidence first; the metric, the harness and the result follow**

발행 CONDERI AI산업전략연구실 | 문서등록번호 CONDERI-AR-2026-11-18 | 문의 contact@conderi.com