MONTHLY EDTECH SOLUTION ANALYSIS REPORT

월간 에듀테크 솔루션 분석 보고서

Vol. 03

2026. 08

이번 호 주제 ISSUE TOPIC

AI 튜터링 플랫폼

대화형 AI 튜터의 교수법 설계와 실증 근거의 범위

대표 분석 솔루션 CS50 Duck (Harvard CS50) · Duolingo Max (Duolingo)

키워드 AI 튜터 · 고등교육 · 교수법 설계 · 정답 노출 · 가상 조교

분석 프레임 4-QUADRANT FRAMEWORK

① 기술 구조	② 시장 동향	③ 교육적 적용	④ 한계·논쟁점
Technology	Market	Pedagogy	Limitations

한림대학교 AI 에듀테크 연구과제

글로컬대학 30 한림대학교 AI 에듀테크센터 연구과제「고등교육 AI 에듀테크의 통합적 발전 방안 연구」[연구 2] 산출물

발간 2026. 8. 31.

EXECUTIVE SUMMARY

핵심 요약

이번 호의 질문

정답을 곧바로 주지 않고 학습을 이끄는 대화형 AI 튜터에서 효과적인 교수법 설계 요소(소크라테스식 질문·스캐폴딩·오개념 진단)는 무엇이며 그 효과는 과목·인지 수준·기간·학습자 특성에 따라 어디까지 일반화되는가? 대학 강좌에 AI 튜터를 넣을 때 교수자·조교의 역할과 운영 모델은 무엇인가?

한 줄 정의

AI 튜터링 플랫폼은 특정 교과·강좌·기능 학습을 위해 교수법을 설계한 학습자 대면 대화형 튜터로, 정답을 곧바로 주기보다 질문·단계적 힌트·오개념 교정·피드백으로 학습을 이끄는 것을 목표로 한다.

핵심 발견

1. 정답을 곧바로 주지 말라는 교수법 지시는 프롬프트만으로 안정되게 지켜지지 않았다. CS50 Duck 의 대화 메시지 1,000 만 건 가운데 약 210 만 건(전체 상호작용의 22%, GPT-4·GPT-4o 기간 누적 자료)의 응답에 코드 블록이 들어 있었고 GPT-4 에서 GPT-4o 로 바꾸자 코드 블록 생성이 예상과 달리 늘었다(2025 년 논문, 개발팀 수행). 벤치마크에서 정답을 노출하지 않은 응답 비율은 7 개 모델 사이에 53.12%에서 94.79%까지 달랐다. 배치된 서비스는 요청 점검(CS50 Duck)과 응답 사후 점검·재작성(CodeHelp)을 덧대고 있고 부분별 순차 안내, 설명·풀이 게이팅, 다중 에이전트 점검 같은 구조는 연구에서 시험되고 있으나 상당수가 시뮬레이션·개발팀 평가 단계다. 오개념 진단은 다단계 문제의 오류 식별과 거짓 경보가 약점이었고 검증기를 붙인 연구 1 편에서 오류 겨냥 정확도가 높아졌으며 여러 연구가 사람의 검토를 권고했다.
2. 하버드 학부 물리 RCT 는 같은 내용의 교실 능동학습 대비 0.73~1.3 SD 의 효과를 보고했지만 이해·적용·분석 수준의 초회 학습에서 튜터 개발팀이 수행한 결과다. 학부생 454 명의 무작위 현장 실험에서는 시험 성취 효과크기가 d = 0.015 였고 독일 입문 프로그래밍 RCT(개발팀 수행)에서는 스캐폴딩 튜터가 실습 점수를 높였지만 지식 향상의 우위는 없었다. 개입 기간의 효과는 챗봇 메타분석마다 방향이 엇갈렸고(포함 연구 대부분은 ChatGPT 확산 이전 챗봇) 무작위 실험에서 튜터를 실제로 쓴 학생이 14.2%(무료 온라인 강좌)에 그친 경우도 있었다. 근거는 과목·인지 수준·기간·학습자 특성을 밝힌 좁은 조건 안에서만 일반화할 수 있다.
3. 두 대표 솔루션은 자료 기준일까지 공개된 문서와 연구로 평가했고 실측이 필요한 항목(게이트 B3·C1 포함 14 개)은 판정하지 않아 미측정으로 두었다. 필수 항목 13 개 가운데 5 개만 판정되어 도입 검토 의견은 두 제품 모두 판정 범위가 부족하다는 뜻의 판정 유보였다. 문서 판정만 담긴 공통 11 개 항목 기준 점수는 CS50 Duck 29.6 점, Duolingo Max 32.4 점이다. 게이트 항목 C5 가 두 제품 모두 부분 충족이고 필수 항목 B9(두 제품)·C4(CS50 Duck)가 미충족이어서 이 판정들이 바뀌지 않는 한 대상 기능을 실측해도 의견의 상한은 조건부 검토다.
4. KAIST 가상 조교 강좌에서 학생의 58%가 인간 교수자에게 질문을 망설인 적이 있다고 답했고 이들은 가상 조교를 더 편하게 느끼며 더 자주 썼다(개발팀 수행, 집단 비교). 그러나 가상 조교에 대한 신뢰는 인간 교수진보다 낮게 유지됐고 국내 대학 사례에서도 응답 검수와 교정 피드백은 사람의 몫으로 남았다. 고등교육법상 학생 교육·지도는 교원의 임무이고 조교의 임무는 교육·연구·학사 사무의 보조이며 AI 기본법 시행령은 2026 년 1 월 22 일부터 고영향·생성형 AI 기반 운용 사실의 사전 고지를 정하고 있다. 한국어 성능 근거는 대부분 튜터가 아닌 LLM 평가였고 모델 업그레이드 뒤에도 언어 간 격차가 일부 문항에서 남았다는 결과(학습자 대상 아님)가 있어 한국어 튜터 품질을 영어 성능으로 짐작하기 어렵다.

4 분면 요약

① 기술 구조

○ 강좌 튜터는 강좌 자료 검색 증강 생성(RAG)을 답의 근거로 쓰고 일부(Jill Watson)는 출처 인용·근거 일치 검증을 더했다. 정답 보류 지시가 지시 희석과 기반 모델 교체로 흔들리기 때문에 배치된 서비스는 요청 점검과 응답 사후 점검·재작성을 덧대고 있으며 순차 안내·다중 에이전트 점검 같은 구조적 통제는 주로 연구 단계에서 시험되고 있다.

○ 오개념 진단은 다단계 문제의 오류 식별과 거짓 경보가 약점이어서 검증기(시험 연구 1 편)·사람 검토와 결합하는 방안이 제안·권고되고 있다. 음성 대화가 학습 이득에서 텍스트보다 낫다는 근거는 없고 음성 인식 오류가 발음 피드백을 왜곡할 수 있으며 강좌 관련 질문 한정·관련성 필터는 오용을 관리하는 장치로 쓰이고 있다.

② 시장 동향

○ 한 시장조사기관은 세계 AI 튜터 시장을 2024 년 16 억 3,000 만 달러로 추정하고 2030 년 79 억 9,000 만 달러를 전망했으나 추정 방법과 고등교육 몫이 공개 요약에서 드러나지 않아 참고치로만 읽어야 하며 학생이 가장 많이 쓰는 도구는 소비자용 범용 제품이었다.

○ 개인 대상 언어 튜터 Duolingo 에서는 2025 년 AI 비용 압박과 2026 년 초 발표된 AI 기능의 하위 등급 이동이 겹쳤다(2026 년 상반기에는 AI 단가 하락으로 매출총이익률 반등). API 운영비가 적게 든 대학 자체 튜터 사례와 공동 이용 시도, K-12 에서 먼저 나타난 공공 조달, 학생의 생성형 AI 채택을 이유로 든 과제 도움 서비스의 구조조정도 나타났다.

③ 교육적 적용

○ LLM 튜터를 쓴 대학 강좌 실험의 결과는 비교 조건·측정·설계에 따라 큰 정적 효과부터 무효과까지 흩어졌다. 큰 효과를 보고한 하버드 물리 실험(이해·적용·분석 수준의 초회 학습)은 개발팀 수행 연구이며 같은 단기 구조화 조건에서 지식 향상 우위가 없던 실험도 있어 인지 수준·기간의 경계는 판단할 수 없다.

○ 학생 질문을 끌어내는 대화 구조에는 초기 근거가 있으나 소크라테스식 질문을 대학 강좌에서 검증한 무작위 실험은 찾지 못했다. 오류 예제를 더한 규칙 기반 튜터 실험 1 편에서는 추가 이득이 없었고 스캐폴딩을 구조로 넣은 튜터의 RCT 두 편(개발팀 수행)은 큰 효과(하버드 물리)와 지식 향상 우위 없음(독일 입문 프로그래밍)으로 갈렸다. AI 튜터는 교수자에게 묻기를 망설이는 학생의 질문 창구가 될 수 있지만 신뢰·교정 피드백과 응답 검수에서는 사람의 역할이 남는다.

④ 한계·논쟁점

○ 효과의 일반화 한계, 교수법적 탈옥과 정답 노출, 학습자가 가려내기 어려운 환각과 교과 내용 오류, 낮은 자발적 사용이 핵심 쟁점이다. 무작위 실험에서 실제 사용률이 14.2%(무료 온라인 강좌)에 그친 경우가 있었고 정답 노출과 교과 오류의 근거는 주로 벤치마크·시뮬레이션에서 나왔다.

○ 대학생의 동반자 목적 AI 사용(17%, 횡단 상관)과 인지적 과의존, 한국어 성능 격차, 학습자 집단별 응답 품질을 직접 잰 고등교육 연구의 부재(형평성), 대학 지침의 책임 개인 전가, 고등교육 챗봇 연구의 철회(2026 년 8 월) 때문에 근거와 운영 양쪽에서 신중함이 필요하다.

대표 솔루션 진단 요약

솔루션	사용성 (A)	교육적 특성(B)	AI 신뢰성 (C)	종합	공통 11 개 항목 기준	도입 검토 의견
CS50 Duck (Harvard CS50)	50.0	26.9	33.3	35.8	29.6	판정 유보
Duolingo Max (Duolingo)	미산출	19.2	50.0	32.4	32.4	판정 유보

주: 두 제품 모두 자료 기준일까지 공개된 문서와 연구로 평가했고 점수는 문서 판정 항목(CS50 Duck 12 개, Duolingo Max 11 개, 필수 각 5 개)에서 나왔다. CS50 Duck 의 축 A 는 1 개 항목 값이고 Duolingo Max 의 종합은 축 B·C 로 재정규화해 두 종합 점수는 직접 비교하지 않는다. 공통 11 개 항목 기준은 두 제품 모두 판정한 문서 판정 항목만의 점수다. 필수 판정 10 개 미만은 판정 유보이며 게이트 C5 가 둘 다 ◐이고 필수 B9(두 제품)·C4(CS50 Duck)가 ○여서 이 판정들이 바뀌지 않는 한 실측해도 상한은 조건부 검토다. 점수는 순위가 아니다.

국내 적용 시사점

○ 대학이 학내 AI 튜터를 학생에게 제공할 때는 화면의 AI 사용 고지, 강좌 자료 연결의 저작권 이용 범위(수업 목적 이용·프로그램 특례·수강생 외 접근 제한), 대화 기록의 위탁·국외 처리·보관 조건을 도입 전에 점검해 둘 만하다(부록 4).

○ 'AI 조교'라는 이름의 서비스는 법률상 조교와 다른 것으로 읽히고 교원의 교육·지도 임무를 AI가 대신할 근거 규정은 조문에서 확인되지 않는다(조문에 기댄 연구진 해석). 이에 AI는 반복 질문과 즉시 응답을, 조교는 응답 검수와 심화 지도를, 교원은 교육·지도와 평가 설계를 맡는 분담을 운영 기준으로 정하는 방안을 검토할 수 있다(분담의 학습 효과는 검증되지 않음).

○ 튜터는 선택 자원으로만 두기보다 수업 활동과 결합해 사용 목적을 안내하고 도입 전 한국어 강좌 자료로 정답 보류와 내용 정확성을 실측한 뒤 AI 없이 치른 평가·지연 측정을 둔 학내 효과 측정으로 확대 여부를 판단할 수 있다.

CONTENTS

목 차

Ⅰ. 배경 및 분석 개요 1

1. 선정 배경 1
2. 분석 범위와 분석 단위 3
3. 분석 대상 솔루션 3
4. 분석 방법 6

Ⅱ. 솔루션군 동향 분석: 4분면 8

1. 기술 구조 8
2. 시장 동향 14
3. 교육적 적용 19
4. 한계·논쟁점 25

Ⅲ. 대표 솔루션 심층 분석 32

1. 솔루션 프로필 32
2. 평가척도 적용 결과 34
3. 비교 종합 38

Ⅳ. 시사점 및 종합 40

1. 국내 적용 시사점 40
2. 본교·연구회 활용 제언 44
3. 종합 46

부록 48

부록 1. 참고 자료 목록 48
부록 2. 평가척도 채점표 55
부록 3. 용어 해설 65
부록 4. 기관 도입 점검표 66

Ⅰ. 배경 및 분석 개요

이 보고서는 「월간 에듀테크 솔루션 분석 보고서」의 셋째 호로 AI 튜터링 플랫폼을 다룬다. Vol.01 이 범용 대규모 언어 모델(LLM)의 학습 모드를, Vol.02 가 학습 경로·난이도 자동 조정을 분석하며 경계로 남긴 교과·강좌 특화 대화형 지도를 이어받는다. 연구과제 사업계획서가 짚은 국제 의제 'LLM 기반 튜터'도 이 호에서 직접 다룬다. 대상 월은 2026 년 8 월이고 자료 기준일은 발간일과 같은 2026 년 8 월 31 일이다. 이 보고서는 기준일까지 공개된 자료만 다루며 기준일 뒤에 발행된 자료와 그에 근거한 내용은 싣지 않는다.

1. 선정 배경

가. 정책 환경

대학생의 생성형 AI 이용은 빠르게 늘었다. OECD(2026b)에 따르면 영국 학부생의 생성형 AI 이용률은 2024 년 초 66%에서 2025 년 92%, 2026 년 95%로 올랐다. 같은 보고서는 고등교육 생성형 AI 이용 근거 다수가 비대표 표본 설문이라 정확한 추정치가 아닌 경향으로 해석해야 한다고 덧붙였다.

범용 도구의 한계도 지적된다. OECD(2026a)의 디지털 교육 전망 보고서는 현재 도구 대부분이 범용으로 설계돼 기성 챗봇이 교육과정과 거의 맞지 않는다고 지적했다. 보고서는 교육 전용 생성형 AI 시스템을 지지하는 주장을 소개하며 이런 도구가 교사에게 도구의 환각 수준을 설정하고 학생의 생성형 AI 상호작용에 피드백을 줄 수 있게 한다고 설명했다.

영국 정부는 2026 년 4 월 안전한 개인 맞춤형 AI 튜터링 도구 개발에 에듀테크 기업과 AI 연구소의 입찰을 요청했다. 이용 대상은 K-12(초·중등 교육) 단계의 취약계층 학생이다. 정부는 9~10 학년 영어·수학·과학·외국어 도구를 학교에서 시험한 뒤 2027 년부터 성공한 도구를 전국에 보급하려 한다(Department for Science, Innovation and Technology & Department for Education, 2026). 그러나 언론이 보도한 시범 계약서(K-12)에서 영국 정부는 시판 AI 튜터링 도구가 수량·범위·근거 기반 모두 제한적이고 완전한 튜터링을 제공하는 제품이 드물다고 평가했다(Turner, 2026).

국내에서는 인공지능기본법(AI 기본법) 시행령이 2026 년 1 월 22 일부터 시행됐다. 시행령 제 23 조제 1 항은 고영향·생성형 AI 를 이용한 제품·서비스를 제공하는 인공지능사업자에게 해당 AI 에 기반해 운용된다는 사실을 이용자에게 사전에 고지하도록 정했다(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령, 2026). 고지 방법과 대학 튜터에 대한 적용 문제는 IV장 1 절 가에서 다룬다.

OECD 와 영국 정부의 문서에서는 정책의 관심이 범용 도구에서 교육 전용 튜터로 옮겨 가는 흐름이 읽힌다. 튜터 보급을 추진하는 영국 정부조차 시판 도구의 근거가 제한적이라고 평가했다. 대학이 튜터를 도입할 때는 근거의 한계와 국내 고지 의무를 감안해 근거가 어디까지 일반화되는지(Q2)와 운영·고지 조건(Q3)을 따로 점검할 필요가 있다.

나. 대학 현장

대학이 강좌 전용 튜터를 직접 만드는 사례가 늘었다. 하버드 CS50 강좌 팀은 CS50 Duck(CS50.ai)을 직접 만들었고 공식 문서는 CS50.ai 가 24 시간 오피스아워와 1:1 에 가까운 학생 대 교사 비율을 지향한다고 밝힌다(CS50, Retrieved 2026). 가상 조교를 직접 운영하는 대학도 있다. 조지아공대(Georgia Tech) 개발팀이 보고한 ChatGPT 기반 Jill Watson 은 별도의 모델 훈련 없이 작동하며 XiaoIce 에서 착안한 스킬 기반 모듈형 설계로 새 API 를 통합할 수 있게 만들어졌다(Taneja et al., 2024).

국내에서는 KAIST 연구팀이 LLM 기반 가상 조교(VTA)를 개발해 대학원생 477 명이 수강한 입문 AI 프로그래밍 강좌에 배치했다고 보고했다(Kweon et al., 2025). 대학 전문지인 한국대학신문도 2025 년 10 월 보도에서 AI 튜터를 도입한 국내 대학이 잇따르고 있다고 전했다(임연서, 2025).

CS50 팀(개발팀 수행)은 2025 년 논문에서 모델이 복잡한 시스템 지시를 지키지 못하는 지시 희석 문제를 보고했다. 이 논문의 GPT-4·GPT-4o 기간 누적 자료에 따르면 학생과 CS50 Duck 의 대화에서 전체 상호작용의 22%에 해당하는 응답에 코드 블록이 들어 있었다(R. Liu et al., 2025). 제공사인 Khan Academy(칸아카데미)는 Khanmigo 에 접근할 수 있는 학생(K-12) 가운데 약 15%만 이용한다고 발표했으나 이용의 정의와 측정 기간은 공개하지 않았다(Khan Academy, 2026). 두 운영 과제의 상세는 Ⅱ장 1 절 나와 Ⅱ장 4 절 라에서 다룬다.

대학은 강좌 전용 튜터를 직접 만들거나 외부 플랫폼으로 운영하는 쪽으로 움직이고 있다. CS50·Jill Watson·KAIST 사례와 국내 대학 보도가 이를 뒷받침한다. 그러나 CS50 팀의 지시 희석 보고와 Khanmigo(K-12)의 낮은 이용률에 비추어 튜터가 교수법 지시를 지키는지와 학습자가 실제로 쓰는지가 운영을 좌우하는 것으로 보인다. 설계 요소(Q1)와 교수자·조교의 운영 모델(Q3)에 관한 질문은 이 흐름에서 나온다.

다. 연구 동향

연구에서는 LLM 튜터에 대한 기대와 과제가 함께 거론된다. 교육 AI 기업 소속 연구자가 참여한 Chu 외(2025)의 서베이는 LLM 에이전트가 복잡한 교수 과업을 자동화하고 교수·학습 과정을 개선하며 교육을 바꾸고 있다고 보았다. 같은 서베이는 LLM 에이전트를 교육 현장에 배치할 때의 핵심 과제로 윤리 문제, 환각과 과의존, 기존 교육 생태계와의 통합을 꼽았다.

튜터로서의 성능에는 의문이 제기된다. 학생 오류 상황의 튜터 응답을 평가한 Maurya 외(2025)는 최신 LLM 이 아직 AI 튜터로서 충분히 좋지 않으며 개선 여지가 크다고 결론지었다(벤치마크 평가). 적대적 학생 에이전트로 튜터의 정답 누출을 시험한 J. Zhao 외(2026)는 LLM 의 기본적 도움 성향이 교수 원리와 자주 충돌한다고 보았다(실제 학생 없음).

효과 근거도 연구마다 방향이 다르다. 하버드 학부 물리 강좌의 무작위 통제 실험(RCT)에서 저자들(개발팀 수행)은 선형회귀 효과크기는 천장효과 때문에 과소추정된 값이라 보고 천장효과를 피하는 분위회귀로 0.73~1.3 표준편차(SD)의 효과크기를 추정했다(Kestin et al., 2025). 반면 튀르키예 고등학교 RCT(K-12)에서는 가드레일 없는 GPT-4 집단(GPT Base)이 AI 없이 치른 시험에서 통제집단보다 성적이 17% 낮았다. 같은 실험에서 교사가 설계한 힌트를 주도록 지시한 GPT Tutor 집단의 시험 성적은 통제집단과 통계적으로 구별되지 않았다(Bastani et al., 2025). 두 실험은 Ⅱ장 3 절 나·다에서 자세히 살핀다.

큰 효과를 보고한 연구와 학습 저해를 보고한 연구가 나란히 쌓이고 같은 실험 안에서도 설계에 따라 결과가 갈렸다. 이 보고서는 쟁점을 어떤 설계가 어떤 조건에서 효과를 내는지로 좁혀 설계 요소, 일반화 범위, 운영 모델의 세 핵심 질문으로 살핀다.

○ Q1. 효과적인 AI 튜터의 교수법 설계 요소(소크라테스식 질문·스캐폴딩·오개념 진단)는 무엇인가?

○ Q2. 실험 근거는 어디까지 일반화할 수 있는가(과목·인지 수준·기간·학습자 특성)?

○ Q3. 대학 강좌에 AI 튜터를 넣을 때 교수자·조교의 역할과 운영 모델은 무엇인가?

2. 분석 범위와 분석 단위

튜터링 시스템은 상호작용의 단위와 튜터 기능으로 나눠 볼 수 있다. VanLehn(2011)은 대부분의 지능형 튜터링 시스템(ITS)이 단계 또는 하위단계 수준으로 상호작용하고 기타 튜터링 시스템은 답 수준 인터페이스를 쓴다고 정리했다. 튜터 기능에 관해서는 Maurya 외(2025)가 학습과학 원리에 기반한 8 개 교수 차원으로 학생 오류 상황에서 LLM 튜터 응답의 교수적 가치를 평가하는 통합 분류 체계를 제안했다. 8 개 차원은 오류 식별, 오류 위치, 정답 노출, 안내 제공, 실행 가능성, 일관성, 튜터 어조, 인간 유사성이다.

이번 호는 특정 교과·강좌·기능 학습을 위해 교수법을 설계한 학습자 대면 대화형 튜터를 분석 단위로 삼는다. 답만 판정하는 도구나 대화 상대에 그치는 범용 비서가 아니라 질문, 단계적 힌트, 오개념 교정, 피드백 같은 튜터 기능을 대화 안에서 수행하도록 설계된 시스템이 대상이다. 범용 어시스턴트의 일반 대화는 Vol.01, 학습 경로·난이도 자동 조정은 Vol.02 에서 다뤘고 계획·점검·성찰을 돕는 메타인지 코칭은 Vol.06 에서 다룬다. 다른 호와의 경계는 [표 Ⅰ-1]과 같다.

범용 LLM 의 학습 모드는 같은 교수법 지시가 전용 튜터와 범용 모델에서 어떻게 다르게 구현되는지 견주기 위해 채점하지 않는 비교 항목으로 포함했다(Ⅱ 장 1 절 라, Ⅱ 장 3 절 나, Ⅳ 장 1 절 라). 진단 → 보완학습 → 재진단형 코스웨어는 경로 자동 조정에 가까워 Vol.02 의 범위로 두고 이번 호에서는 다루지 않는다.

[표 Ⅰ-1] 분석 범위와 경계

구분	이번 호에서 다루는 것	제외·제한(넘기는 호)
사용 주체	학습자(대화로 지도받는 학습자)	교수자용 자료 생성(Vol.07), 채점·피드백 생성(Vol.05)
핵심 기능	교과·강좌 특화 대화형 지도(질문·단계적 힌트·오개념 교정·피드백)	학습 경로·난이도 자동 조정(Vol.02), 계획·점검·성찰 코칭(Vol.06)
튜터 유형	대학 강좌 전용 튜터·가상 조교, 언어 학습 앱의 AI 대화 기능, 범용 LLM 학습 모드(비교 항목)	범용 어시스턴트의 일반 대화(Vol.01), 진단 → 보완학습형 코스웨어(Vol.02)
대화 방식	텍스트·음성 대화, 강좌 자료 기반 응답	정서 동반·감정 인식 기능은 위험 쟁점으로만(Ⅱ 장 4 절 마·사)
학교급	고등교육	K-12 제품·연구는 경계·참고 사례로 'K-12' 표시
근거 범위	독립 연구(효과), 제공사 문서(제품 사실), 동료심사 전 자료('동료심사 전'), 개발팀 수행 연구('개발팀 수행'), 제공사 수행·후원 연구('자체 보고')	제공사 마케팅 성과 수치는 쓰지 않음, 본교(한림대학교) 자체 서비스는 평가하지 않음(Ⅳ 장 2 절 가)

자료: 연구진 정리

3. 분석 대상 솔루션

대표 솔루션은 국내외 균형, 시의성, 사업 연속성, 국내 도입 실적, 국내 이용 가능성, 실사용 진단 가능성의 6 개 기준으로 골랐다. 두 제품 모두 6 개 기준을 다 충족하지는 못했으나 설계(Q1)와 운영 모델(Q3)의 유형 대비를 위해 대학이 직접 만든 강좌 전용 튜터와 개인 소비자 대상(B2C) 언어 학습 앱의 AI 튜터 기능에서 하나씩 골라 CS50 Duck 과 Duolingo 의 Duolingo Max 를 채점 대상으로 삼았다

. 특히 Duolingo Max는 대학 단위 도입 실적과 Max 기능 단위의 독립 효과 근거가 없는 상태에서 대학 강좌 튜터와 대비되는 개인 구독형 언어 튜터의 설계·운영 모델을 보려고 골랐다. 동향 분석 대상과 국내 사례를 포함한 전체 대상은 [표 Ⅰ-2]와 같다.

CS50 Duck의 공식 문서에 따르면 CS50.ai는 cs50.ai 웹과 CS50용 Visual Studio Code(cs50.dev)에 함께 들어 있는 학생·교사용 ChatGPT 변형이며 답을 곧바로 주기보다 학생을 답으로 이끄는 좋은 튜터처럼 동작하도록 설계되었다(CS50, Retrieved 2026). 시의성과 사업 연속성 면에서 2026년 가을 하버드 CS50 실라버스는 과제에서 CS50 Duck 등 CS50 자체 AI만 허용하고 ChatGPT·Claude·Copilot·Gemini 등 다른 AI 소프트웨어는 금지한다(CS50, 2026).

이용 조건 면에서 CS50 Duck은 CS50 강좌가 과제에 허용하는 자체 AI로 쓰인다(Ⅲ장 1절). 국내 이용 가능성 면에서는 외부 기관이 CS50 Duck을 자기 강좌에 들이는 경로를 공개 문서로 확인하지 못했고 국내 대학의 도입 실적도 찾지 못했다. 제공사와 독립된 근거로는 뮌헨응용과학대 연구팀의 혼합연구(통제집단 없음)가 있는데 이 연구는 자교 컴퓨팅 사고(Computational Thinking) 강좌에서 CS50 Duck에 대한 학생 인식·참여·실제 활용을 살폈다(Zönnchen et al., 2025).

Duolingo Max는 2023년 3월 출시 게시본에서 Duolingo의 유료 구독 등급 Super보다 위의 등급으로 Explain My Answer와 Roleplay를 제공했다. 같은 글의 현행 본문은 Video Call과 Roleplay를 생성형 AI 기술로 동작하는 두 기능으로 안내한다(Duolingo, 2023). 시의성 면에서 Duolingo는 2026년 1월 게시글에서 구독자 전용이던 Explain My Answer를 모든 학습자에게 무료로 제공한다고 밝혔다(Mas Castillo, 2026).

이용 조건 면에서 Max 기능은 iOS·Android 앱에서만 쓸 수 있고 웹에서는 구매할 수 없다(Duolingo, Retrieved 2026-a). 사업 연속성 면에서 Duolingo는 학교용 채널인 Duolingo for Schools를 2027년 7월 31일 종료할 예정이라고 공지했다(Duolingo, 2026f). 국내 이용 가능성 면에서는 한국어 도움말과 약관 문서를 제공하지만 Explain My Answer·Roleplay 제공 과정 목록에 한국어 사용자용 영어 과정은 없다(Ⅱ장 4절 바). 국내 대학의 도입 실적도 찾지 못했다. 제공사와 독립된 고등교육 연구로는 Duolingo Max를 다른 앱·AI 도구와 묶어 처치한 연구 2편을 찾았으나 Max 기능의 효과를 따로 가를 수 없고 단일집단 설계나 낮은 측정 신뢰도 같은 연구 질의 한계가 있어 효과 근거에서 제외했다.

동향 분석 대상은 채점하지 않고 Ⅱ장의 4분면 분석에서 다룬다. 대학이 직접 개발한 가상 조교인 KAIST VTA와 Jill Watson은 Ⅰ장 1절 나에서 소개했다. Speak는 2019년 12월 출시된 인공지능 기반 영어 회화 학습 앱이다(공정거래위원회, 2024). Khan Academy의 Khanmigo는 K-12 중심이어서 경계 사례로, 과제 도움 서비스 Chegg는 시장 축소의 참고 사례로 다룬다. 범용 LLM의 학습 모드는 Vol.01의 분석 대상이어서 이번 호에서는 채점하지 않는 비교 항목으로 둔다. 연구에서 이런 모드를 평가할 때는 근사 프롬프트를 쓰기도 한다. S. Zhao 외(2026)는 Google의 LearnLM과 OpenAI의 ChatGPT 학습 모드 같은 일부 튜터형 시스템이 API로 공개되지 않아 LearnLM 시스템 지시문과 학습 모드를 흉내 낸다고 알려진 유포 프롬프트를 다듬어 기준 튜터링 프롬프트로 썼다.

국내 솔루션은 채점하지 않았다. 국내 대학이 자체 개발한 KAIST 가상 조교와 국내 한 대학의 파이썬 입문 강좌에 배치된 SCALA는 수강생 계정이 있어야 쓸 수 있고 단국대 AI 페르소나는 교수법 설계를 공개 자료로 확인하지 못해 외부에서 실사용 진단을 하기 어렵다. 엘리스(Elice)의 AI 헬피는 기관 관리자가 기능을 활성화한 경우에만 기관 구성원에게 제공된다(엘리스, Retrieved 2026). 충북대 AI-Tutor는 공개 자료로 대화형 튜터인지 확인하지 못해 사례 서술에서 뺐다. 대표 솔루션은 모두 국외 제품이어서 국내외 균형 기준은 충족하지 못했고 국내 후보는 공개 접근과 이해관계 문제로 채점에서 빼 사례(Ⅰ장 1절 나, Ⅱ장 2절 다, Ⅳ장 2절 가)로 다룬다.

이 보고서는 HAI-GPT를 운영하는 본교 AI에듀테크센터의 연구과제 산출물이며 연구진에는 해당 부서·센터 구성원이 포함된다. 이 센터가 운영·실증하는 HAI-MathTutor(AI 선형대수 튜터)·HAI-Tutor·K-University 'AI 동반자'와 HAI-GPT 스터디 튜터는 이해관계 때문에 채점하지 않고 국내 사례로만 다룬다(IV장 2절 가). 이 보고서가 게시되는 KELI 플랫폼의 솔루션 마켓에는 센터가 운영하는 HAI-MathTutor(솔루션 공급 기업 슈퍼러닝)가 함께 소개되어 있다.

자료 수집·분석, 초안 작성, 채점 초안 작성에 보조로 활용한 생성형 AI는 Anthropic의 Claude다. Anthropic은 범용 LLM 학습 모드를 제공하는 사업자이며 두 대표 솔루션에 기반 모델을 공급하거나 데이터 처리를 맡는 OpenAI·Microsoft·Google과 범용 LLM 시장에서 경쟁 관계에 있다. 한편 영국 경쟁시장청 결정문에 따르면 Google과 Anthropic의 파트너십에는 Google의 Anthropic 무의결권 주식 연속 취득과 Google의 Anthropic 컴퓨팅 용량 비독점 공급 약정이 포함된다(Competition and Markets Authority, 2024). Anthropic은 2025년 10월 최대 100만 개의 AI 칩(TPU)을 포함해 Google Cloud 기술 사용을 확대할 계획이라고 발표했다(Anthropic, 2025b). Microsoft도 2025년 11월 Anthropic에 최대 50억 달러를 투자하기로 했고 Anthropic은 Azure 컴퓨팅 용량 300억 달러 구매를 약정했다(Microsoft Corporate Blogs, 2025).

CS50 개발팀의 2025년 논문에 따르면 대표 솔루션인 CS50 Duck은 학생 메시지를 Azure에 호스팅된 GPT-4o로 중계하는 CS50.ai 서버로 동작한다(R. Liu et al., 2025). CS50 문서는 CS50 Duck을 지원해 준 GitHub·Microsoft·OpenAI에 감사를 표한다(CS50, Retrieved 2026).

Duolingo는 Video Call 등 AI 기능에서 제출한 텍스트·오디오를 OpenAI·Google 같은 AI 공급업체와 공유할 수 있다고 밝힌다(Duolingo, 2026c). Anthropic은 2025년 7월 Claude for Education이 학습관리시스템 Canvas의 LTI를 지원해 학생이 Canvas 강좌 안에서 Claude를 바로 쓸 수 있게 되었다고 발표했다(Anthropic, 2025a). 이 연동은 범용 LLM의 기관용 기능이어서 이번 호에서는 채점하지 않는 비교 항목(범용 LLM 학습 모드)으로 본다. 분석 대상 후보에 Anthropic의 강좌 전용 튜터가 없어 이 관계로 채점에서 뺀 제품은 없다. 전용 튜터와 범용 LLM 학습 모드의 비교는 II장 1절 라, II장 3절 나, IV장 1절 라에서 다루고 두 대표 솔루션의 판정 근거는 부록 2에 공개한다.

[표 Ⅰ-2] 분석 대상 솔루션 개요

구분	솔루션(제공사)	튜터 유형	최근 변화·근거	이번 호 처리
대표	CS50 Duck(Harvard CS50)	대학 강좌 전용 튜터(강의 자료 기반 응답·교수법 지시, 웹·VS Code)	2026년 가을 강좌의 자체 AI 한정 정책, 지시 희석·코드 블록 응답 보고(2025년 논문, 개발팀 수행), 독립 혼합연구 1편(학생 인식·효과크기 없음)	채점(문서 판정)
대표	Duolingo Max(Duolingo)	언어 학습 앱의 생성형 AI 대화 기능(Roleplay·Video Call)	Explain My Answer 무료화(2026. 1.), Video Call의 Super 이동·Max 구성 실험 발표(2026. 2.), 신규 Super 구독자 대부분에 Video Call 제공(2026. 8. 주주서한), Duolingo for Schools 종료 예정(2027. 7. 31.), Max 기능 단위 독립 효과 연구 없음	채점(문서 판정)

구분	솔루션(제공사)	튜터 유형	최근 변화·근거	이번 호 처리
동향	KAIST VTA	대학원 강좌 가상 조교(강좌 관련 질문 응답)	2024 년 가을 477 명 강좌 배치·사용 평가(개발팀 수행)	동향 분석
동향	Jill Watson(Georgia Tech)	온라인 대학원 강좌 가상 조교	ChatGPT 기반 재구축, 실재감 RCT(2023 년 가을·개발팀 수행)	동향 분석
동향	Speak	AI 기반 영어 회화 학습 앱	시리즈 C 투자(2024. 12.), 국내 환불 제한 약관 시정(2024. 12.)	동향 분석
경계	Khanmigo(Khan Academy)	K-12 튜터	개편 시범·2026 년 여름 배포 계획(제공사 발표), 군집 RCT(K-12·동료심사 전)	경계 사례(K-12)
참고	Chegg	과제 도움 서비스	구조조정·학술 서비스 매출 감소(2025)	참고
비교	범용 LLM 학습 모드(OpenAI·Google·Anthropic 등)	범용 LLM 의 교수법 모드	일부 API 미공개, 연구에서는 근사 프롬프트로 평가	비교 항목(Vol.01)
사례	SCALA(국내 한 대학 파이썬 입문 강좌)	강좌 튜터(예상 질문 사전 생성·선제적 지원)	2025 년 봄 학기 배치·개발팀 평가(기업 공저자 참여)	국내 사례
사례	단국대 악셀(AX-EL)·에듀아이(EduAI)·AI 페르소나	교수 지식을 학습한 AI 페르소나·학습지원 플랫폼	2026. 7. 보도	국내 사례
사례	엘리스 AI 헬피	코딩 실습 튜터(교육자 지도서 기반)	교육자 매뉴얼(기관 관리자 활성화)	국내 사례
사례	충북대 AI-Tutor	—	공개 자료로 대화형 여부 미확인	이름만
사례	본교 HAI-MathTutor(AI 선형대수 튜터)·HAI-Tutor·K-University 'AI 동반자', HAI-GPT 스터디 튜터	교과목 전용 튜터·학습지원 서비스·교수 업무 지원 조교	2025 년 가을 도입(AI 선형대수 튜터, 본교 발표), 2026. 6. 서비스 매칭데이(본교 발표)	IV장 2 절 가 사례(평가하지 않음)

자료: 각 사 공식 문서·발표, 독립 연구, 언론 보도를 바탕으로 연구진 정리

4. 분석 방법

분석은 [표 Ⅰ-3]의 다섯 단계로 진행했다. 자료를 수집·선별하고 기술 구조, 시장 동향, 교육적 적용, 한계·논쟁점의 4 분면으로 분석한 뒤 효과 근거를 등급화했다. 이어 대표 솔루션 두 개에 평가척도 채점을 적용하고 국내 적용 시사점을 해석했다. DOI 가 있는 논문은 OpenAlex 와 Crossref 로 철회·정정 여부를 확인했다. 효과 근거는 설계·대상·표본·효과크기를 기록해 근거 수준을 나누고 K-12, 동료심사 전, 개발팀 수행, 자체 보고 여부를 표시했다. 개발팀 수행은 평가 대상 튜터를 연구자가 직접 개발·운영한 연구이고 자체 보고는 제공사가 수행하거나 후원한 연구다. CS50 팀의 논문처럼 개발팀이 곧 제공사인 경우 개발팀 수행은 제공사 자체 보고이기도 하다. 평가척도의 효과성 증거 항목(B9·B10)에는 제공사와 독립적인 고등교육 연구만 반영했다.

이번 호의 평가는 자료 기준일까지 공개된 문서(제공사 도움말·정책·공시와 연구)에 근거하며 실사용 진단 결과는 싣지 않았다. 자료 수집·분석, 초안 작성, 채점 초안 작성에는 생성형 AI 를 보조로 활용했고 판정은 연구진이 근거 문서와 대조해 확정한다.

평가척도 v1 은 사용성(A, 8 개), 교육적 특성(B, 10 개), AI 신뢰성·윤리·거버넌스(C, 8 개)의 26 개 항목이다. 항목 판정(● 2 점, ◐ 1 점, ○ 0 점)에 가중(필수 ×3, 권장 ×2, 보조 ×1)을 곱해 축마다 100 점으로 환산하고 판정하지 않은 항목(미측정)과 해당 없는 항목(N/A)은 계산에서 뺀다. 측정 항목이 없는 축은 점수를 내지 않는다. 종합 점수에는 이번 호 가중치 A:B:C = 30:40:30 을 적용하되 점수가 있는 축의 가중치로 다시 맞추고 두 제품의 비교에는 공통 측정 항목(11 개)의 점수만 쓴다. 문서 판정(DOC) 항목은 두 제품 모두 자료 기준일까지 공개된 문서 근거로만 판정했다. 실측 관찰이 있어야 판정하는 AUTO·HYBRID 항목은 이번 호에 실측 결과를 싣지 않아 모두 미측정이다.

도입 검토 의견은 게이트 항목(B3 피드백·적응, C1 타당성·신뢰성과 환각 관리, C5 인간 통제·교수자 역할)과 종합 점수로 정한다. 적극 검토는 게이트 모두 ●·필수 항목에 ○ 없음·종합 80 이상, 조건부 검토는 게이트 모두 ◐ 이상·종합 60 이상이며 그 밖은 신중 검토다. 필수 항목 13 개 가운데 측정한 항목이 10 개 미만이면 경계값을 적용하지 않고 '판정 유보'로 적는다. 고부담 게이트는 이번 호에 해당하지 않는다. 축 점수의 반올림 규칙은 Ⅲ장 2 절 다에 밝힌다.

[표 Ⅰ-3] 분석 방법 요약

단계	내용	자료·도구·조건	산출
① 자료 수집·선별	학술·정책·제공사 문서와 언론 보도 수집, 철회·정정 확인	OpenAlex·Crossref	부록 1
② 4 분면 분석	기술 구조, 시장 동향, 교육적 적용, 한계·논쟁점	수집 자료	[표 Ⅱ-1]~[표 Ⅱ-4]
③ 효과 근거 등급화	설계·대상·표본·효과크기·근거 수준 기록, 'K-12·동료심사 전·개발팀 수행·자체 보고' 표시	메타분석·실험·준실험	[표 Ⅱ-3]
④ 평가척도 채점	26 개 항목 판정·가중(실측 필요 항목은 미측정), 게이트 B3·C1·C5, 고부담 게이트 해당 없음	평가척도 v1, A:B:C = 30:40:30, 자료 기준일까지 공개된 문서·연구	[표 Ⅲ-3], 부록 2
⑤ 국내 적용 해석	법령·공공 안내서·국내 사례 대조	법령 원문, 정부 안내서, 국내 연구·보도	[표 Ⅳ-1]·[표 Ⅳ-2], 부록 4

자료: 연구진 작성

Ⅱ. 솔루션군 동향 분석: 4 분면

Ⅱ장은 Ⅰ장 4 절의 분석 절차에 따라 AI 튜터링 플랫폼을 네 분면으로 살핀다. 기술 구조(Ⅱ장 1 절)는 핵심 질문 Q1 의 설계 요소가 실제 시스템에서 어떻게 구현되는지 보고 시장 동향(Ⅱ장 2 절)은 공급 구조와 대학 도입 경로에서 Q3 의 운영 모델 선택지를 찾는다. 교육적 적용(Ⅱ장 3 절)은 Q2 에 맞춰 효과 근거의 일반화 범위를 점검하며 한계·논쟁점(Ⅱ장 4 절)은 남은 쟁점을 모은다. 근거의 성격(K-12, 동료심사 전, 개발팀 수행, 자체 보고, 시뮬레이션)은 문장 안 괄호로 밝히고 Ⅱ장 1·3 절 소결에는 Q1·Q2 의 잠정 답을, 2 절 소결에는 Q3 의 운영 모델 선택지를, 4 절 소결에는 쟁점이 Q2·Q3 에 주는 함의를 적는다.

1. 기술 구조

가. 강좌 자료 기반 응답과 근거 제시

강좌 튜터가 먼저 풀어야 할 과제는 답의 근거다. 교육 분야의 검색 증강 생성(RAG) 연구를 정리한 체계적 서베이에 따르면 LLM 은 환각과 고정된 내부 지식 때문에 교육 현장에서 신뢰성이 제한된다. RAG 는 외부 지식 기반에서 관련 정보를 검색해 LLM 생성 과정에 넣는 방식이다. 저자들은 RAG 가 사실 정확도를 높이고 지식의 동적 갱신을 가능하게 해 LLM 을 교육 응용에 적합하게 만든다고 설명했다(Li et al., 2025).

대학 강좌 튜터는 이 방식을 강좌 자료에 적용한다. 2024 년 개발팀 논문에 따르면 CS50 Duck 이 동작하는 서비스인 CS50.ai 는 LLM 을 미세조정하거나 재학습하지 않고 강의 자막 기반 RAG 로 강의 관련 질문 답변의 환각 가능성을 줄였다(R. Liu et al., 2024). CS50 팀은 2025 년 논문에서 미세조정 결과도 보고했다(Ⅱ장 1 절 나). 조지아공대 개발팀의 논문에 따르면 Jill Watson 은 답의 출처 문서를 인용하고 텍스트 함의 판정(검색한 문맥이 답을 논리적으로 뒷받침하는지 가리는 방법)으로 근거 일치를 검증해 환각에 대응한다. 검색한 문맥이 생성한 답을 완전히 함의하지 못하면 신뢰도가 낮다는 경고를 답 앞에 붙이고 사용자가 직접 확인하도록 권한다(Taneja et al., 2024).

국내 상용 서비스도 교육자가 만든 자료를 답의 근거로 삼는다. 코딩 실습 플랫폼 엘리스의 교육자 매뉴얼(제공사 설명)에 따르면 AI 헬피는 교육자가 작성한 수업 자료 지도서와 지시사항을 기반으로 답변을 생성한다. 학생 코드 역시 교육자가 작성한 설명·지시사항·해설 코드에 맞춰 분석하고 안내한다(엘리스, Retrieved 2026).

세 사례 모두 강좌 자료를 답의 근거로 삼고 그중 Jill Watson 은 출처 인용과 근거 일치 검증까지 더했다. 같은 강좌 자료 기반 응답이라도 무엇을 근거로 답했는지 학습자에게 보여 주는 정도는 제품마다 다르다. 이번 호가 확인한 문서에서 출처 인용과 저신뢰 경고를 학습자에게 보여 준다고 밝힌 사례는 Jill Watson 이다. 강좌 자료를 튜터에 연결할 때 따르는 저작권·자료 이용 범위 문제는 IV장 1 절 가에서, 환각 완화와 검색 지식의 완전성 같은 남은 기술 과제는 Ⅱ장 4 절 다에서 다룬다.

나. 교수법 지시와 그 준수 한계

정답을 바로 주지 않는 교수법은 일차적으로 시스템 프롬프트의 지시로 구현된다. CS50 개발팀의 2024 년 논문에 따르면 CS50.ai 는 정답을 직접 주지 않고 학생을 안내한다는 교수 철학에 맞춘 '교수법적 가드레일'을 내장했다. GPT-4 요청마다 강좌별 규칙을 담은 시스템 프롬프트와 학생 질의를 담은 사용자 프롬프트를 같이 보내는 방식이다(R. Liu et al., 2024).

현장 보고에 따르면 이런 지시가 늘 그대로 지켜지지는 않는다. 프로그래밍 수업용 도우미 CodeHelp 개발팀은 LLM 이 한 프롬프트에서 정확히 따를 수 있는 지시의 수와 복잡도에 한계가 있어 여러 목표를 단일 프롬프트로 일관되게 달성할 수 없었다고 보고했다. 개발팀이 사용한 LLM 은 프롬프트가 달리 요구해도 완전한 해답을 주는 쪽으로 강하게 치우친 것으로 보였다(Liffiton et al., 2023). CS50 팀은 2025 년 개발팀 논문에서 복잡한 시스템 지시를 모델이 지키지 못하는 문제를 지시 희석이라 부르며 GPT-4 에서 GPT-4o 로 바꾸자 코드 블록 생성이 예상과 달리 늘었다고 보고했다. 학생과 CS50 Duck 의 대화 메시지 1,000 만 건을 분석한 결과 약 210 만 건의 응답(전체 상호작용의 22%, GPT-4·GPT-4o 기간 누적 자료)에 코드 블록이 들어 있었다(R. Liu et al., 2025). 현재 코드 블록 비율은 공개 자료에서 확인되지 않았다.

같은 경향은 연구용 평가에서 확인된다. 시뮬레이션 학생과의 다회차 대화로 튜터를 훈련·평가한 연구는 직접 질의응답에 최적화된 LLM 이 답을 전략적으로 보류해야 하는 효과적 교수법을 자주 훼손한다는 문제의식에서 출발했다. 실험 결과는 공들인 프롬프트로도 표준 LLM 이 가르치기보다 답하기에 최적화돼 있다는 가설을 지지했다(Dinucu-Jianu et al., 2025). 학생 오류가 담긴 수학 대화에서 7 개 LLM 의 튜터 응답을 사람이 평가한 결과 GPT-4 는 응답의 약 47%에서 정답을 노출해 실행 가능성이 떨어졌다. 이 수학 오류 교정 대화 벤치마크에서 정답을 노출하지 않은 응답의 비율은 7 개 LLM(GPT-4·Gemini·Sonnet·Mistral·Llama-3.1-405B·Llama-3.1-8B·Phi3) 사이에 53.12%에서 94.79%까지 모델에 따라 크게 달랐다(Maurya et al., 2025). 이 수치는 연구용 대화 자료에서 잰 값이어서 상용 튜터의 실제 정답 노출률로 읽을 수는 없다.

고등학교 수학 다단계 문제로 만든 튜터 벤치마크(K-12 수학 내용, 실제 학생 없음)에서도 현행 LLM 들은 수학 지식이 뛰어나고 문제 풀이에 강해도 스캐폴딩(비계) 응답 생성 과제에서는 일관되게 부족했다. SocraticLM·LearnLM·Meta 의 Llama 등 여러 모델은 프롬프트로 교수법 지시를 주어도 성능이 떨어지거나 비슷했다. 교수법 행동을 열거하는 프롬프트는 작성이 번거롭고 작은 변화에 민감하며 시험하기 어렵다는 지적도 있었다(Macina et al., 2025).

학생이 일부러 정답을 끌어내려 할 때는 위험이 더 커진다. GSM8K 등 수학 문항을 중심으로 공개 모델을 튜터로 쓴 적대적 학생 에이전트 시뮬레이션에서 맥락 조작 기법은 모델 평균 74%의 노출률로 가장 높은 정답 누출을 유도했다. 모델별 평균 누출률은 40%에서 75%(Alibaba 의 기본 모델 Qwen-7B 와 교수법 정렬 튜터 TutorRL-7B)까지였다. 같은 논문의 확장 실험에서 OpenAI 의 GPT-5 는 정답 누출 응답이 4.58%로 소형 공개 모델보다 크게 낮았지만 정답을 숨겨도 지나치게 많은 정보를 주는 경향이 있었다. 코딩 영역으로 넓힌 실험에서는 프로그래밍 같은 절차적 영역에서 중간 단계 스캐폴딩이 완전한 해답 노출로 이어지기 쉬웠다(J. Zhao et al., 2026). CS50 개발팀도 GPT-4 를 속여 숙제 정답 전체를 받아내려는 프롬프트 주입 공격을 막기 위해 학생 요청을 점검하는 '가드' 기능을 두었다(R. Liu et al., 2024).

실제 서비스는 지시 위에 사후 통제를 덧대고 있다. CodeHelp 는 주 응답을 두 번 생성해 코드 블록과 교수자 지정 금지어 포함 여부로 채점하고 더 나은 쪽을 고른다. 최고 점수 응답에도 코드가 남으면 세 번째 프롬프트로 응답을 다시 써서 코드를 없앤다(Liffiton et al., 2023). 모델 자체를 손보는 시도도 있다. CS50 팀의 후속 연구(2025 년 개발팀 논문)에서 저자들은 소규모 고품질 데이터셋으로 미세조정한 결과 다회차 대화 평가에서 교수 목표와의 정렬이 크게 개선됐다고 보고했다(R. Liu et al., 2025).

정답 보류는 CodeHelp·CS50 팀의 보고와 벤치마크 결과에서 보듯 프롬프트 지시만으로 안정되지 않으며 같은 지시라도 기반 모델이 바뀌면 지켜지는 정도가 달라질 수 있다. 실제 서비스와 연구가

요청 점검, 응답 사후 점검·재작성, 미세조정 같은 층을 덧대는 배경도 여기에 있는 것으로 보인다. 프로그래밍처럼 풀이 절차가 뚜렷한 교과일수록 정답 노출 위험을 더 엄격히 살펴야 한다. 대학이 튜터를 도입하기 전에 정답 직접 요구와 적대적 요청에 튜터가 어떻게 응답하는지 직접 확인하기를 권한다. 이번 호에는 두 대표 솔루션의 이런 응답에 대한 실측 결과를 싣지 않아 게이트 항목 B3(피드백·적응)·C1(타당성·신뢰성과 환각 관리)을 미측정으로 두었다(Ⅲ장 2절).

다. 오개념 진단과 오류 피드백

오개념 진단은 학생 응답에서 오류의 원인을 짚어 내는 기능이다. 한 연구는 미국 조지아주 기술대학 체계의 대학 대수 수업(2024년 봄)에서 나온 지능형 튜터 기록으로 GPT-4의 학생 오류 진단을 평가했다. 이 연구에서 저자들은 GPT-4가 대학 대수의 학생 오류를 약 80% 정확도로 진단했다고 개괄했다. 문제 풀이의 어느 단계에서 작업 중인지 설명하는 맥락을 충분히 주자 정확도는 71.9%에서 87.8%로 올랐다. 그러나 확장된 맥락을 주어도 여러 단계로 된 복잡한 문제의 오류 식별에서는 성능이 낮았다(Reddig et al., 2025).

진단 다음 단계인 피드백의 질은 고르지 않았다. 같은 연구에서 GPT-4가 만든 피드백의 34.2%는 지나치게 일반적이거나 틀렸거나 정답을 그대로 알려 주는 힌트였다. 저자들은 낮은 품질의 응답을 걸러 내려면 현재 사람의 감독이 필요하다고 밝혔다(Reddig et al., 2025). 오픈소스 적응형 튜터 OATutor에서는 Mechanical Turk의 성인 학습자(대학생 표본 아님) 274명을 대상으로 무작위 배정 온라인 실험이 진행됐다. 이 실험에서 ChatGPT 힌트 조건만 무도움 통제 대비 유의한 학습 향상을 보였고 인간 튜터 힌트와의 차이는 유의하지 않았다. 품질 점검에서는 ChatGPT 생성 힌트가 문제의 32%에서 실패했다. 자기일관성 기법을 적용하자 품질 실패율은 대수 문제에서 거의 0%, 통계 문제에서 13%로 줄었다(Pardos & Bhandari, 2024).

대학 강좌 로그로 만든 벤치마크도 비슷한 약점을 드러냈다. 이 연구는 미국 대형 대학 학부 이산수학 강좌에 배치된 명제논리 튜터의 2023년 봄 상호작용 로그에서 증명 상태 516개를 추출해 LLM 피드백 에이전트 7종을 평가했다. 에이전트들은 최적 단계에서는 천장에 가까운 성능을 보인 반면 타당하나 비최적인 추론은 체계적으로 과잉 기각하고 틀린 풀이는 과잉 승인했다. 다만 평가한 풀이는 실제 학생이 아니라 각 모델이 학생 역할로 생성한 것이다(LLM 시뮬레이션 벤치마크). 저자들은 지식그래프 기반 모델이 진단을 맡고 LLM은 열린 스캐폴딩과 대화를 맡는 혼합 구조를 제안했다(Yasir et al., 2026).

K-12 자료에서는 맞힌 답 뒤에 숨은 오개념이 문제로 떠올랐다. 영국 중등 수학 플랫폼의 학생 응답(K-12 데이터)을 분석한 연구는 학생이 잘못된 추론으로 정답에 이르면 모델이 오개념을 놓치는 '정답 함정' 실패 유형을 보고했다. 이 오개념 탐지에서 최신 LLM은 84%, Google의 T5를 미세조정한 모델은 57%의 정확도를 보였다. 가장 성능이 좋은 모델도 실제 탐지 1건당 약 4건의 거짓 경보를 내 단독 선별에 쓰기 어려웠다. 저자들은 전체 정확도가 높아도 결정적 실패가 가려질 수 있어 학생 추론 분석에는 여전히 사람의 판단이 필요하다고 결론지었다(Imran & Bulathwela, 2026).

다단계 수학 문장제(K-12 수준)로 실험한 연구는 오류를 찾는 일 자체가 어렵다는 결과를 내놓았다. 이 연구에서 기존 LLM은 추론 문제는 잘 풀지만 학생 오류를 정확히 찾아내 피드백을 그 오류에 맞추는 데는 어려움을 겪었다. 참조 풀이 없이 프롬프트만 쓴 모델은 특히 틀린 풀이를 식별하는 성능이 낮았다. 연구팀은 학생 풀이 검증을 응답 생성과 분리한 모듈형 구조를 제안했다. 자동·인간 평가에서 검증기를 붙이자 기존 기준 모델보다 학생 오류를 더 정확히 겨냥하고 더 자주 옳으며 환각이 적은

응답이 나왔다. 검증으로 응답이 더 표적화되는 대신 실행 가능성은 떨어지는 경우도 있었다(Daheim et al., 2024).

국내 연구(중학교 수학 문항, K-12)에서는 오개념 지식을 진단 문항으로 옮기는 단계의 한계가 드러났다. 이 연구는 GPT-4o 로 오류 유형 5 종과 일차방정식 문항 유형 5 종에 따라 오개념 기반 오답지 후보 321 개를 생성했다. 전체 시도 중 정답과 다른 오답지를 만든 비율은 약 27.7%였고 추상적인 개념상 오류 유형은 오답지를 전혀 만들지 못했다. 저자들은 LLM 이 오개념 규칙을 기계적으로 적용할 뿐 교육적 맥락에서 타당한지는 고려하지 못한다고 해석했다(조상연 외, 2025).

대학 자료와 K-12 자료에 걸쳐 LLM 단독의 자동 판정에는 다단계 문제의 오류 식별과 거짓 경보라는 약점이 거듭 나타났다. 결합 방안으로는 검증기를 붙인 연구 1 편에서 오류 겨냥 정확도가 높아졌고 지식그래프 결합은 제안 단계이며 여러 연구가 사람의 검토를 권고했다. 이에 튜터를 실측할 때는 오개념 교정과 이해 점검 과업을 넣고 거짓 경보를 걸러 낼 사람의 검토 절차를 마련해야 한다. 정확한 진단이 실행 가능한 피드백으로 이어지는지는 별개의 문제이며 Ⅱ장 4 절 다에서 살핀다.

라. 구조 수준의 교수법 설계

Ⅱ장 1 절 나·다에서 본 한계에 대응해 교수법을 프롬프트 문구가 아니라 시스템 구조에 넣으려는 시도가 나오고 있다. 대학 선형대수 벤치마크 연구(실제 학생 없음)는 학생이 정답 유도 프롬프트로 해답을 끌어내는 '교수법적 탈옥'을 다뤘다. 저자들은 유능한 범용 모델에 '좋은 교사가 되라'고 프롬프트하는 것만으로는 부족하며 교수법 원리를 구조 수준에 넣어야 한다고 결론지었다. 그 방안으로 질의에서 선수 개념을 추론하고 숙달 격차를 찾아 설명과 문제 풀이 사이를 명시적으로 게이팅하는 그래프 증강 튜터링 파이프라인을 제안했다(S. Zhao et al., 2026). 튜터 개발팀이 수행한 하버드 물리 RCT(Ⅱ장 3 절 나)의 AI 튜터는 능동적 참여 촉진·인지부하 관리·성장 마인드셋 촉진 지침을 시스템 프롬프트에 담았다. 그러나 시스템 프롬프트만으로는 여러 부분으로 된 문제를 안정적으로 스캐폴딩할 수 없어 플랫폼이 각 문제를 부분별로 순서대로 안내하도록 구조화했다. 정확하고 질 높은 설명을 위해 프롬프트에 단계별 완전 풀이도 넣었다(Kestin et al., 2025).

시뮬레이션 평가에서는 모델을 교수 목표에 맞게 다시 훈련하거나 여러 에이전트가 응답을 점검하게 하는 방법을 시험하고 있다. 매개변수 70 억 개(7B) 규모의 튜터를 인간 주석 없이 시뮬레이션 학생과의 온라인 강화학습으로 훈련하자 LearnLM 같은 대형 독점 모델과 비슷한 성능에 이르렀다. 이 연구는 보상 가중치로 교수법적 지원과 학생 해결률 사이의 균형을 조절할 수 있음을 보였다(Dinucu-Jianu et al., 2025). 다만 Ⅱ장 1 절 나의 적대적 학생 공격 시뮬레이션에서 교수법 정렬 튜터 TutorRL-7B 는 기본 모델 Qwen-7B 와 함께 평균 누출률 75%로 가장 취약했고 기본 모델보다 더 많은 대화 턴을 거친 뒤 정답을 드러냈을 뿐이다(J. Zhao et al., 2026). 같은 시뮬레이션의 사전 정의 공격에서 다중 에이전트 튜터와 추론 튜터 방어 전략은 Qwen-32B 의 누출률을 46%에서 각각 4%와 2%로 낮췄고 저자들은 이를 가장 큰 감소로 밝혔다. 적대적 학생 역할로 미세조정한 에이전트의 공격에서는 두 방어 전략이 Qwen-7B 튜터의 누출을 82%에서 각각 61%와 38%로 줄였으나 충분히 완화하지는 못했다. 확장 실험에서는 Qwen2.5-72B 튜터가 Qwen-32B 보다 누출이 많아 모델 규모가 견고성을 보장하지 않았다(J. Zhao et al., 2026).

학생이 먼저 묻기를 기다리지 않는 선제적 지원도 구조 설계에 속한다. 국내 한 대학의 파이썬 입문 강좌에 배치된 SCALA(개발팀 수행, 기업 공저자 참여)는 강의 전에 학생이 할 법한 질문과 답을 미리 생성하는 예측 질의 관리 기능을 도입했다. 학생이 먼저 질문해야 응답하는 기존 반응형 AI 조교와 달리 선제적 지원을 하도록 설계됐으며 전체 세션의 51.2%가 사전 생성 질문에서 시작됐다(J. Lee et

al., 2026). Khan Academy 가 일부 학군에서 시범 운영한 개편 Khanmigo(K-12, 제공사 발표)는 학생이 완벽한 질문을 하기를 기다리지 않고 필요한 순간에 도움을 제안한다. 도움 요청 연구를 근거로 문제를 시도하기 전과 후의 지원 방식을 나누었고 기능 숙달도와 선수 기능 정보를 받아 필요하면 선수 기능 복습을 제안한다(Khan Academy, 2026).

Ⅲ장 1 절 가~라의 사례에서는 Q1 의 설계 요소를 프롬프트 문구보다 부분별 순차 안내, 설명·풀이 게이팅, 다중 에이전트 점검, 정렬 학습, 선제 질문 같은 시스템 구조로 구현하려는 시도가 이어지고 있으나 교수법 정렬 튜터도 적대적 요청에는 취약했다. 범용 LLM 의 학습 모드는 외부에서 내부 구조를 확인하기 어려워 이런 비교에서는 근사 평가에 그친다(Ⅰ 장 3 절). 범용 모델에 교수법 지시를 얹는 방식은 공들인 프롬프트로도 답하기 쪽으로 치우치고 스캐폴딩 응답 생성에 약했으며(Ⅱ 장 1 절 나) 유능한 범용 모델에 좋은 교사가 되라고 지시하는 것만으로는 부족하다는 결론도 나왔다. 대학이 튜터를 고를 때는 어떤 지시를 넣었는지보다 정답 보류와 스캐폴딩을 어떤 구조로 강제하는지가 선택 기준이다. 이 구조들의 상당수가 아직 시뮬레이션이나 개발팀 평가 단계에 있다는 점도 고려해야 한다.

마. 음성 대화와 운영 장치

음성 대화 기능은 외국어 튜터에서 개발·평가되고 있다. 한국전자통신연구원(ETRI) 연구팀은 초등학생용 영어 튜터 EBS AI 펭톡(K-12)과 외국인용 한국어 튜터를 개발해 직접 평가했다. 두 튜터는 정밀한 전사를 바탕으로 발음과 말하기 유창성을 평가하고 오류 피드백을 제공한다. 연구팀은 비원어민 발화는 오류가 잦아 고성능 자유발화 음성 인식이 필요하다고 밝혔다(Kang et al., 2024).

음성 인식의 오류는 발음 피드백을 왜곡할 수 있다. Speak 소속 저자가 공개 제 2 언어(L2) 영어 말뭉치로 수행한 기술 평가에서 최신 음성 인식의 언어모델은 실제 발음보다 의도한 단어를 복원해 발음 오류를 체계적으로 가린다는 분석이 나왔다. 같은 연구는 재학습이나 모델 내부 접근 없이 음소 수준 음향 유사도로 인식 후보를 다시 고르는 방법을 제시했다. 이 방법은 잘못된 발음을 맞는 것으로 받아들이는 오발음 오수용률을 6.0%p(L2-ARCTIC)와 5.6%p(speechocean762) 낮췄다(Park, 2026).

제공사 블로그의 설명에 따르면 Duolingo Max 의 대화 기능은 사람이 쓴 시나리오에 모델 생성을 더한다. Roleplay 시나리오와 첫 메시지는 Duolingo 전문가가 쓰고 학습자의 과정 위치에 맞추며 모델에 대화가 나아갈 방향을 지시한다. Video Call 의 캐릭터 릴리는 실시간으로 대화를 이어 가고 이전 통화 내용을 기억한다(Duolingo, 2023). 제공사는 Explain My Answer 가 학습자가 범한 오류와 해당 규칙이 학습자에게 어려운 이유를 고려해 개인화 피드백을 만든다고 설명한다(Mas Castillo, 2026). 세 기능 모두 제공사의 기능 설명이며 학습 효과를 보여 주는 자료는 아니다.

음성·영상 튜터에 감정 추론 기능을 붙이면 규제 경계를 따져야 한다. EU 집행위원회의 금지 관행 지침(비구속, 교육기관 적용 범위는 Ⅱ장 4 절 사)은 타자 방식·표정·자세·동작에서 감정을 추론하는 AI 를 생체정보 기반이어서 금지 범위에 드는 것으로 본다(European Commission, 2025).

응답 범위를 관리하는 운영 장치도 튜터 구조의 일부다. CodeHelp 는 강좌와 관련된 질문에만 답해 무제한 LLM 접근 통로로 악용되는 것을 막는다. 학습 도구 상호운용성(LTI) 표준을 지원하는 Moodle·Canvas 같은 학습관리시스템(LMS)과도 연동된다(Liffiton et al., 2023). Jill Watson 은 질문 관련성 분류기, 유해 텍스트 필터, 정중한 응답을 유도하는 프롬프트를 안전 장치로 쓴다(Taneja et al., 2024).

음성 대화가 학습 이득에서 텍스트보다 낫다는 근거는 아직 없고 출판 편향을 보정한 챗봇 메타분석에서는 텍스트 기반의 효과크기가 더 컸다(Ⅱ 장 3 절 라). 음성 인식 오류가 교정 피드백을

왜곡할 위험도 남아 있다. 강좌 관련 질문 한정과 관련성 분류·유해 필터는 오용을 관리하는 장치로 쓰이고 있다. 음성 튜터를 도입할 때는 발음 피드백의 정확성을 점검하고 감정 추론 기능이 있다면 규제 경계부터 확인해야 한다. Ⅱ장 1절에서 살핀 구성 요소를 구현 방식, 사례, 확인된 한계로 정리하면 [표 Ⅱ-1]과 같다.

[표 Ⅱ-1] AI 튜터의 기술 구조 비교

구성 요소	구현 방식	사례	확인된 한계
근거 제시	강좌 자료 기반 RAG, 출처 인용, 텍스트 함의 기반 근거 일치 검증, 저신뢰 경고	CS50 Duck(강의 자막), Jill Watson(텍스트 함의 검증), 엘리스 AI 헬피(교육자 지도서, 제공사 설명)	제한된 문맥으로 답해야 하는 한계, 검색 지식의 완전성(Ⅱ장 4절 다)
교수법 지시	시스템 프롬프트(정답 보류·강좌별 규칙·학습 지침)	CS50 Duck, CodeHelp, 하버드 물리 튜터	단일 프롬프트의 지시 수·복잡도 한계, 지시 희석(2025년 논문의 GPT-4·GPT-4o 기간 응답 22%에 코드 블록), 맥락 조작 요청의 정답 누출(공개 모델 튜터 평균 74%, 시뮬레이션), 정답 미노출 응답 53.12~94.79%(7개 모델, 연구용 벤치마크로 상용 튜터 수치 아님)
응답·요청 통제	응답 2회 생성·채점, 코드 제거 재작성(응답), 요청 점검 '가드'(입력)	CodeHelp, CS50 Duck	통제 효과를 비교 측정한 공개 자료 미확인(CS50 Duck 현재 코드 응답 비율 미확인), 절차적 영역의 해답 노출
오개념 진단	LLM 오류 진단, 풀이 검증기 분리, 지식그래프 결합(제안), 사람 검토	대학 대수 튜터 기록 평가, 명제논리 튜터 벤치마크	다단계 문제의 오류 식별 저조, 피드백 34.2% 저품질, 정답 함정·거짓 경보(탐지 1건당 약 4건, K-12 데이터)
구조 수준 교수법	부분별 순차 안내, 설명·풀이 게이팅, 다중 에이전트·추론 튜터 방어, 강화학습 정렬, 소규모 데이터셋 미세조정	하버드 물리 튜터, 그래프 증강 파이프라인, 7B 정렬 튜터, 다중 에이전트·추론 방어 실험(사전 정의 공격의 최대 감소 46% → 4%·2%, 시뮬레이션), CS50 Duck(미세조정, 개발팀 수행)	다수가 실제 학생 검증 전(시뮬레이션·벤치마크), 교수법 정렬 튜터의 적대적 요청 취약성(TutorRL-7B 평균 누출률 75%, 시뮬레이션), 방어 효과의 모델·공격별 차이(Qwen-7B 82% → 61%·38%, 시뮬레이션)
선제적 지원	예상 질문 사전 생성, 필요한 순간의 도움 제안, 선수 기능 복습 제안	SCALA(국내 한 대학), Khanmigo 개편(K-12·제공사 발표)	효과 근거 제한적(Ⅱ장 3절 마)
음성·역할극	음성 인식·발음 평가, 전문가 작성 역할극 시나리오, 이전 통화를 기억하는 영상 통화 캐릭터	Duolingo Max(Roleplay·Video Call, 제공사 설명), ETRI 언어 튜터(K-12·외국인 한국어 학습자), Speak(소속 저자의 음성 인식 평가)	음성 인식의 발음 오류 은폐, 텍스트 대비 학습 이득 근거 없음(Ⅱ장 3절 라), 감정 추론 기능의 규제 경계(EU 지침)

구성 요소	구현 방식	사례	확인된 한계
운영 장치	강좌 관련 질문 한정, 관련성 분류·유해 필터, LMS 연동(LTI)	CodeHelp, Jill Watson	안전 장치가 답할 수 있는 질문까지 거르는 상충(Ⅱ장 4절 다)

자료: Ⅱ장 1절 인용 자료를 바탕으로 연구진 정리

주: 시뮬레이션·벤치마크는 실제 학생 없이 평가한 연구다. Duolingo Max·Khanmigo·엘리스 AI 헬피의 기능은 제공사 설명이며 효과 근거가 아니다. 개발팀이 수행한 평가(CS50 Duck·CodeHelp·Jill Watson·하버드 물리 튜터·SCALA·ETRI 언어 튜터)는 본문에 표시했다. ETRI 언어 튜터는 초등학생용 영어 튜터(K-12)와 외국인 한국어 학습자용 튜터다.

소결 Ⅱ장 1절의 근거로는 현재 구현에서 확인되는 Q1의 설계 요소가 네 갈래다(학습 효과는 Ⅱ장 3절 라에서 점검). 구체적으로는 강좌 자료에 근거해 답하는 장치(일부는 출처 인용·근거 일치 검증까지), 정답 보류 지시를 요청·응답 점검과 순차 안내 같은 구조로 강제하는 장치, 검증기나 사람 검토와의 결합이 제안된 오개념 진단, 학생의 질문을 끌어내거나 먼저 도움을 제안하는 상호작용이다. 설명·풀이 게이팅은 이 가운데 둘째 요소에 속하지만 벤치마크에서 제안된 단계다. 프롬프트에만 기댄 설계가 기반 모델 교체와 적대적 요청에 흔들린다는 점은 개발팀 보고와 벤치마크·시뮬레이션에서 거듭 확인됐다. 적대적 요청에는 프롬프트형 튜터뿐 아니라 교수법 정렬 튜터도 취약했고 시뮬레이션에서는 다중 에이전트·추론 방어가 누출을 크게 줄였으나 감소 폭은 모델과 공격 방식에 따라 달랐다. 구조 수준 설계의 근거 상당수는 시뮬레이션이나 개발팀 보고에서 나온 것이어서 실제 학생의 학습 효과로 이어지는지는 따로 따져 보아야 한다. 정답 노출 쟁점은 Ⅱ장 4절 나에서 다루며 Q1의 최종 답은 Ⅳ장 3절에서 종합한다.

2. 시장 동향

AI 튜터의 공급 구조를 시장 규모 추정, 개인 대상 언어 튜터, 대학이 직접 만드는 튜터, 기관 라이선스와 공공 조달, 과제 도움 사업자의 순서로 살핀다. 주요 지표의 수치·기준 시점·출처·정의는 [표 Ⅱ-2]에 모았다.

가. 시장 규모 추정의 한계

이번 호가 확인한 AI 튜터 시장 규모 추정은 한 시장조사기관의 공개 요약이다. 시장조사기관 Grand View Research(2025)는 세계 AI 튜터 시장을 2024년 16억 3,000만 달러로 추정하고 2030년 79억 9,000만 달러(연평균 30.5%)를 전망했으며 2025년 시장은 21억 1,000만 달러에 이를 것으로 내다봤다. 이 보고서는 최종 사용자 가운데 K-12가 2024년 최대 매출 비중을 차지했다고 보았다.

학생이 가장 많이 쓰는 도구는 교육용 제품이 아니라 범용 제품이다. 디지털교육위원회(DEC)의 글로벌 학생 조사(2024)를 재인용한 OECD(2026b)에 따르면 가장 많이 쓰인 도구는 ChatGPT(66%)·Grammarly(25%)·Microsoft Copilot(25%)였고 모두 소비자용 범용 제품이었다. 같은 보고서는 이런 이용 조사 다수가 비대표 표본에 기대고 있어 경향으로 읽어야 한다고 밝혔다(Ⅰ장 1절 가).

시장 추정의 공개 요약은 추정 방법과 고등교육 몫을 밝히지 않아 이 보고서는 시장 규모를 참고치로만 제시한다. 학생의 실제 이용이 범용 도구에 쏠려 있다는 점까지 고려하면 대학의 도입 판단에는 시장 성장 전망보다 강좌 단위의 수요와 이미 쓰이는 범용 도구라는 대체재가 더 현실적인 기준이 된다.

나. 언어 학습 B2C 튜터의 성장과 등급 재편

언어 학습 분야에서는 B2C 튜터 앱이 유료 구독과 투자 유치로 시장을 넓혔다. 공정거래위원회(2024)에 따르면 AI 영어 회화 앱 Speak 는 2022 년 국내 주요 앱마켓 교육 분야 매출 1 위였고 2024 년 누적 다운로드 수가 500 만 이상이었다. 같은 자료에서 Speak 는 이용 기간에 따라 월간·연간·평생 3 가지 유료 구독을 제공하며 2024 년 12 월 기본가 기준 이용권은 월간 29,000 원, 연간 129,000 원, 평생 450,000 원이었다. 언론 보도에 따르면 같은 달 Speak 는 Accel 주도로 7,800 만 달러 시리즈 C 를 유치해 기업가치 10 억 달러에 이르렀고 6 개월 전 시리즈 B 연장 투자 때의 기업가치는 5 억 달러였다. Speak 는 OpenAI 기술로 플랫폼을 운영하며 활성 이용자 수는 공개하지 않는다(Lunden, 2024).

연차보고서(2025 회계연도)에 따르면 Duolingo Max 는 2023 년 출시된 상위 구독 등급으로 일부 이용자에게 Super 보다 높은 가격에 제공된다. 같은 보고서에서 전체 매출총이익률은 2024 년 72.8%에서 2025 년 72.2%로 낮아졌고 회사는 Video Call 같은 기능의 AI 비용 증가를 반영한 구독 매출총이익률 하락을 주된 원인의 하나로 들었다. 2025 년 4 월 AI 활용 메모 공개가 부정적 여론과 이용자 증가 둔화에 기여했을 수 있다는 점도 위험 요인에 적었다(Duolingo, 2026b). 2026 년 6 월 30 일 기준 분기보고서에서는 상반기 전체 매출총이익률이 전년 같은 기간 71.8%에서 72.8%로 올랐고 회사는 제 3 자 AI 단가의 지속적 하락을 반영한 구독 매출총이익률 상승을 주된 원인으로 들었다(Duolingo, 2026e).

Duolingo 는 2026 년 2 월 주주서한에서 Video Call 을 Super 등급으로 옮기고 Duolingo Max 의 구성 변경을 실험하겠다고 밝혔다. 2026 년 예약매출(bookings) 성장률은 약 11%로 제시하면서 기존 방식이면 20% 가까이 가능했다고 보았고 수익화 마찰을 줄이는 대가로 5,000 만 달러 이상의 예약매출을 포기하고 무료 이용자 경험에 투자한다고 추산했다. 같은 서한에 따르면 일간 활성 이용자(DAU) 성장률은 2025 년 내내 둔화했다(Duolingo, 2026a). 2026 년 8 월 주주서한은 2 분기 DAU 가 1 년 전보다 23% 늘었다고 밝혔으나 이 지표는 독립된 제 3 자의 검증을 받지 않은 회사 내부 집계다(Duolingo, 2026d).

공공 부문이 개발해 보급한 언어 튜터도 있다. ETRI 연구팀인 Kang 외(2024)는 초등학생용 영어 튜터 EBS AI 펭톡(K-12)과 외국인용 한국어 튜터인 세종학당 한국어 AI 튜터를 개발했고 두 시스템은 국내 정부 기관을 통해 배포됐다고 밝혔다(개발팀 수행).

Duolingo 에서는 AI 기능을 최상위 등급에 묶어 두던 구성을 하위 등급으로 내리는 재편(2026 년 초 발표)과 2025 년 AI 비용 증가에 따른 수익성 압박이 겹쳤고 2026 년 상반기 공시에서는 AI 단가 하락으로 매출총이익률이 다시 올랐다. Speak 는 투자 규모와 기업가치가 보도됐지만 활성 이용자 수를 공개하지 않아 이용 규모를 확인할 공개 지표가 부족하다. 대학이 개인 구독형 튜터를 강좌에 쓰려면 등급 구성과 가격이 바뀔 수 있다는 점을 계획의 전제로 삼아야 한다(IV 장 1 절 라). Explain My Answer 의 무료 제공을 포함한 Duolingo Max 의 구성 변화는 Ⅲ장 1 절에서 자세히 다룬다.

다. 대학이 직접 만드는 강좌 튜터

하버드 CS50 Duck(Ⅰ장 1 절 나) 외에도 대학이 강좌용 튜터를 직접 만든 사례가 있고 그중 일부는 다른 대학과 나누려 한다. ETH 취리히의 가상 조교 Ethel 은 RAG 로 강좌별 참고 자료에 근거해 답하며 대형 기초 물리 강좌에서 1,800 명 이상에게 손글씨 과제 피드백과 서술형 시험 채점에 쓰였다. ETH 는 다른 대학도 쓸 수 있게 Ethel Commons 를 개발하고 있다. UC 버클리 연구실이 만든 OATutor 는 고등교육용 오픈소스 지능형 튜터링 시스템이다(OECD, 2026b).

대학 자체 튜터 가운데 운영비를 공개한 사례가 있다. KAIST 가상 조교(Ⅰ장 1절 나, 개발팀 수행)의 14주 운영 비용은 API 사용료와 대화 기록 저장을 포함해 약 180달러였다(Kweon et al., 2025). 국내 다른 대학의 AI 튜터 도입도 보도됐다. 영남대는 2025년 9월 AI 기반 맞춤형 학습 지원 챗봇 AI 튜터를 도입했고 세종사이버대 AI 튜터 서비스는 과목별 교안·강의 자막·수업계획서·참고문헌 등을 학습해 학습자 질문에 실시간으로 답한다. 서울사이버대 AI 학습 튜터는 강의 관련 데이터를 학습해 상담·답변을 제공하며 다국어 소통을 지원한다. 기사에 인용된 전문가는 대학 AI 튜터가 ChatGPT·Gemini 같은 범용 서비스와 차별점을 가져야 한다고 말했다(임연서, 2025).

단국대는 AI 기반 교육모델 악셀(AX-EL)을 대표 혁신 사례로 내세운다. 대학 홍보 성격의 기획 기사에 따르면 악셀은 AI·메타버스·확장현실(XR)·액티브 러닝을 접목한 학생 참여형 수업 모델이다. 교수의 지식과 강의 경험을 학습한 AI 페르소나가 실제 교수처럼 강의를 진행하고 학생 질문에 답하는 것이 특징이다. 단국대는 AI 학습지원 플랫폼 에듀아이(EduAI), 학생용 AI 비서 단아이(Dan.i), 교수용 교육분석 시스템 데스크(D-ESK)도 도입했다(김가현, 2026).

범용 모델에 강좌 자료를 얹는 대학 자체 튜터에서 API 운영비가 적게 든 사례는 있으나(KAIST 대학원 강좌 1개, 14주 약 180달러, 개발팀 수행) 이 금액은 API 사용료와 대화 기록 저장 비용이어서 한 사례로 구축 비용을 일반화할 수는 없다. 오픈소스와 공동 이용으로 다른 대학에 넓히려는 시도도 나타난다. 국내 사례 대부분은 교수법 설계와 효과 자료가 공개되지 않은 보도 수준이며 단국대 AI 페르소나의 교수법 설계와 가드레일도 기사에서 확인되지 않는다. 자체 개발 경로는 운영비보다 개발·검수 인력(Ⅳ장 1절 라)과 데이터 조건(Ⅳ장 1절 나)에 좌우될 가능성이 크다.

라. 기관 라이선스와 공공 조달

상용 플랫폼을 기관 단위로 들이는 경로는 LMS 연동과 기관 규정에서 막히기도 한다. Gedera와 Griffiths(2026)는 뉴질랜드 오클랜드공과대(AUT) 경영학 강좌에 AI 조교를 통합한 과정을 분석했다. 연구 대상인 NF AI(Noodle Factory)는 싱가포르 기업이 만든 교육 플랫폼이다. 이 사례에서 LMS인 Canvas와의 연동은 한 강좌에서만 지원됐고 기관의 기술 규정 정비가 필요해 추가 연동이 중단됐다. 국내에도 엘리스 AI 헬피처럼 기관 관리자가 기능을 활성화해야 구성원이 쓸 수 있는 기관형 제품이 있다(Ⅰ장 3절).

공공 조달은 학교 대상 AI 튜터의 새 구매 경로가 되고 있다(K-12). 영국 정부는 AI 튜터링 도구 공모(Ⅰ장 1절 가)에서 선정 기업이 30만 파운드를 받고 최대 8곳이 'Pioneer Group'을 구성한다고 밝혔다. 정부는 연간 최대 45만 명의 취약계층 학생이 혜택을 볼 수 있다고 했다(Department for Science, Innovation and Technology & Department for Education, 2026). 이후 언론 보도에 따르면 시범 운영 계약 8건이 기술 기업들에 돌아갔고 각 공급사는 부가세 포함 36만 파운드를 받는다. 시범 계약이 최종 도구 공급을 보장하지는 않는다(Turner, 2026).

학교 채널을 대하는 전략은 제공사마다 다르다. Khan Academy는 개편한 Khanmigo(K-12, 제공사 발표)를 2026년 여름 모든 학군 파트너에 배포한다고 밝혔다(Khan Academy, 2026). 반면 Duolingo는 Duolingo for Schools 종료(Ⅰ장 3절)를 핵심 학습 경험 개선에 집중하기 위한 결정이라고 설명했고 신규 Duolingo for Schools 계정과 교실은 더 이상 만들 수 없다고 안내했다(Duolingo, 2026f).

학교 단위 구매가 공공 조달과 학군 단위 배포로 넓어지는 흐름이 보이지만 사례는 K-12에 몰려 있다. 대학이 상용 플랫폼을 기관 단위로 들일 때는 LMS 연동과 기관 규정이 걸림돌이 될 수 있고 개인 구독 중심 제공사가 오히려 기관 채널을 줄이는 상반된 움직임도 있다. 기관 라이선스 계약 전 점검은 LMS 연동 범위와 공급사 지원 체계, 채널의 지속성을 대상으로 한다(부록 4).

마. 과제 도움 사업자의 축소

범용 생성형 AI 의 영향은 과제 도움 서비스 기업의 공시에서 드러난다. Chegg 는 2025 년 5 월과 10 월 당시 인력의 약 56%인 약 640 명 감원을 포함한 구조조정 계획을 발표하면서 경쟁 심화와 학생들의 생성형 AI 채택을 이유로 들었다. 연차보고서에는 학생들이 ChatGPT 같은 생성형 AI 를 교육 전문 서비스의 강력한 대안으로 본다고 적었다. 2025 년 학술 서비스 매출은 전년보다 2 억 3,540 만 달러(43%) 줄었다. 회사는 트래픽 감소로 구독자가 줄어 생긴 구독 매출 감소(2 억 2,380 만 달러)를 주원인으로 설명했고 2025 년 말 직원 수는 595 명이었다(Chegg, 2026).

교육 플랫폼은 아니지만 프로그래밍 질문·답변 플랫폼에서도 비슷한 변화가 관찰됐다. 대조 플랫폼과 비교한 준실험 연구인 del Rio-Chanona 외(2024)에 따르면 ChatGPT 공개 후 6 개월 안에 Stack Overflow 활동은 ChatGPT 접근이 제한된 러시아·중국의 유사 플랫폼과 ChatGPT 성능이 낮은 수학 포럼에 비해 25% 줄었다. 저자들은 이 추정치를 ChatGPT 영향의 하한으로 해석했다. 동료 평가로 측정한 게시물 품질에는 유의한 변화가 없었고 콘텐츠 생성 감소는 경험이 많은 사용자와 적은 사용자에서 비슷하게 나타났다. LLM 은 중복·저품질·초급 콘텐츠만 대체한 것이 아니었다.

Chegg 공시와 Stack Overflow 연구에 따르면 범용 생성형 AI 는 질문·답변형 과제 도움 서비스의 수요를 흡수하고 있는 것으로 보인다. 교수법 설계 없이 답을 주는 서비스는 범용 도구와 구별되기 어려워지는 반면 강좌에 결합된 튜터에는 차별점이 생길 여지가 있다(연구진 해석). 대학 튜터가 범용 도구와 같은 답 제공에 머물면 대체재와 차이가 드러나지 않고 반대로 정답을 보류하면 학생이 더 편한 범용 도구로 옮겨 갈 위험이 있다(편안함의 함정, Ⅱ 장 3 절 나). 어느 쪽이든 수업 활동과의 결합이 사용을 좌우할 수 있다(Ⅱ 장 4 절 라).

[표 Ⅱ-2] 시장 지표(수치·기준 시점·출처·정의)

지표	수치	기준 시점	출처	성격	정의·주의
세계 AI 튜터 시장(추정)	2024 년 16 억 3,000 만 달러, 2025 년 21 억 1,000 만 달러 전망, 2030 년 79 억 9,000 만 달러(연평균 30.5%)	2024 년 기준	Grand View Research(2025)	시장조사기관 추정	최종 사용자 중 K-12 가 최대 매출 비중, 고등교육 몫 미공개
학생이 가장 많이 쓴 AI 도구	ChatGPT 66%, Grammarly 25%, Microsoft Copilot 25%	2024	OECD(2026b)	국제기구 문서(DEC 조사 재인용)	모두 소비자용 범용 제품, 비대표 표본 조사의 한계
Speak 국내 지위	국내 주요 앱마켓 교육 분야 매출 1 위, 누적 다운로드 500 만 이상	2022(매출)·2024(다운로드)	공정거래위원회(2024)	정부 보도자료	환불 약관 시정 보도자료 기재
Speak 투자	시리즈 C 7,800 만 달러, 기업가치 10 억 달러(6 개월 전 5 억 달러)	2024. 12.	Lunden(2024)	언론 보도	활성 이용자 수 비공개
Duolingo 매출총이익률(연간)	72.8%(2024) → 72.2%(2025)	2025 회계연도	Duolingo(2026b)	법정 공시(연차보고서)	Video Call 같은 기능의 AI 비용 증가 등을 원인으로 든 회사 설명

지표	수치	기준 시점	출처	성격	정의·주의
Duolingo 매출총이익률(상반기)	71.8%(2025 년 상반기) → 72.8%(2026 년 상반기)	2026. 6. 30. 분기	Duolingo(2026e)	법정 공시(분기보고서)	제 3 자 AI 단가 하락을 원인으로 든 회사 설명
Duolingo 예약매출 전망	2026 년 성장률 약 11%(기존 방식이면 20% 가까이), 5,000 만 달러 이상 포기 추산	2026. 2.	Duolingo(2026a)	제공사 발표(주주서한)	수익화 마찰 축소와 무료 이용자 경험 투자의 대가
대학 자체 가상 조교 운영비	14 주 약 180 달러	2024 년 가을 학기	Kweon 외(2025)	동료심사 논문(개발팀 수행)	KAIST 대학원 강좌 1 개, API 사용료·대화 기록 저장 포함(개발·인력 비용은 이 수치에 없음)
영국 AI 튜터 시범 계약	8 건, 공급사별 36 만 파운드(부가세 포함)	2026. 6. 보도	Turner(2026)	언론 보도	K-12, 최종 도구 공급 보장 아님
Chegg 학술 서비스 매출	2 억 3,540 만 달러(43%) 감소, 인력 약 56%(약 640 명) 감원 포함 구조조정	2025 회계연도	Chegg(2026)	법정 공시(연차보고서)	경쟁 심화와 학생의 생성형 AI 채택을 이유로 듦
Stack Overflow 활동	비교 플랫폼 대비 25% 감소(영향의 하한)	ChatGPT 공개 후 6 개월	del Rio-Chanona 외(2024)	동료심사 준실험	교육 플랫폼 아님(유추 근거)

자료: 각 기관 보고서·공시, 정부 문서, 언론 보도, 학술 연구를 바탕으로 연구진 정리

주: 시장 추정치는 한 조사 기관의 추정이며 정의와 방법이 공개 요약에 드러나지 않아 참고치로만 읽는다. 제공사 발표·공시 수치는 성격 열에 출처 유형을 밝혔다. 영국 시범 계약 금액(36 만 파운드)은 부가세 포함 보도 기준이고 정부 공고의 선정 기업 지원액(30 만 파운드)은 부가세 포함 여부가 밝혀져 있지 않아 두 금액을 직접 비교하지 않는다.

소결 AI 튜터 시장의 규모 추정은 한 조사 기관의 공개 요약에 기대고 추정 방법과 고등교육 몫이 드러나지 않아 참고치로만 읽어야 한다. 개인 대상 언어 튜터는 구독과 투자로 성장했으나 Duolingo 에서는 AI 기능의 하위 등급 이동과 2025 년 AI 비용 부담이 겹쳤고 2026 년 상반기에는 AI 단가 하락으로 매출총이익률이 다시 올랐다. 강좌 자료에 근거해 답하는 대학 자체 튜터는 API 운영비가 적게 든 사례가 있고 오픈소스와 공동 이용으로 다른 대학과 나누려는 시도도 나타났지만 국내 사례는 대부분 설계와 효과 자료가 없는 보도 수준이다. 공공 조달과 학군 단위 배포는 K-12 에서 먼저 나타났고 조달을 추진한 영국 정부도 시판 도구의 근거가 제한적이라고 평가했다(Ⅰ장 1절 가). 대학의 상용 플랫폼 도입은 LMS 연동에서 막히기도 했고 답을 주는 과제 도움 서비스에서는 학생의 생성형 AI 채택을 이유로 한 구조조정이 있었다. Q3 에 비추어 보면 대학의 운영 모델 선택지는 자체 개발·상용 플랫폼·개인 구독으로 나뉘며 그 성패는 각각 개발·검수 인력, LMS 연동·기관 규정, 등급·가격 변동에 걸린 것으로 보인다(Ⅳ장 1절 라).

3. 교육적 적용

가. ITS와 챗봇 메타분석의 효과 범위

AI 튜터를 가늠할 때는 LLM 이전 ITS의 효과를 기준점으로 삼는다. 인간 튜터링·컴퓨터 튜터링·튜터링 없음을 비교한 실험들(K-12·대학 혼합)의 종합 리뷰에서 인간 튜터링의 효과크기는 통념(d = 2.0)보다 훨씬 낮은 d = 0.79였고 ITS는 0.76으로 거의 같은 수준이었다(VanLehn, 2011). 학교급이 섞인 ITS 통제 평가 50건의 효과 중앙값은 0.66 표준편차였다(Kulik & Fletcher, 2016).

효과크기 107개(참여자 1만 4,321명)의 전 학교급 메타분석에서 ITS는 교사 주도 대집단 수업(g = 0.42)·비ITS 컴퓨터 기반 수업(g = 0.57)·교재 또는 워크북(g = 0.35)보다 높은 성취와 관련되었다. 개별 인간 튜터링(g = –0.11)·소집단 수업(g = 0.05)과는 유의한 차이가 없었다(Ma et al., 2014). 대학생 대상 39개 연구(ITS 22종)를 종합한 메타분석에서 ITS는 중간 수준의 정적 효과(g = 0.32~0.37)를 보였고 인간 튜터링보다 덜 효과적이었지만 그 밖의 수업 방법·학습 활동보다는 효과가 컸다. 교수자와 교수법이 중요하다는 근거도 일부 있었다(Steenbergen-Hu & Cooper, 2014).

챗봇 메타분석의 값은 편차가 크다. 학교급이 섞인 한 메타분석에서 AI 챗봇의 전체 학습 성과 효과크기(ES)는 0.964, 고등교육 학습자 집단의 효과크기는 ES = 1.079로 크고 유의했다. 다만 문헌 검색 시점이 2023년 1월 4일이어서 포함 연구 대부분이 ChatGPT 확산 이전 챗봇이다(Wu & Yu, 2024). 출판 편향을 선택 모형으로 보정한 다른 메타분석에서는 전체 효과크기가 g = 0.76에서 0.381로 줄었지만 유의했다(Laun & Wolff, 2025).

학교급이 섞인 보고서 28편(효과크기 31개)의 메타분석은 챗봇의 학습 효과를 g = 0.48의 중간 효과로 보고했고 비교 집단의 수업 유형·실험 기간·챗봇 유형·챗봇 과업이 유의한 조절변수였다(Alemdağ, 2025). 언어 학습 챗봇 연구 28편(효과크기 70개)을 종합한 메타분석에서도 비챗봇 조건 대비 효과는 g = 0.484였고 교육 단계·언어 수준·인터페이스 설계·상호작용 능력이 효과를 조절했다(Wang et al., 2025).

챗봇 메타분석의 전체 효과는 포함 연구와 보정 여부에 따라 0.381에서 0.964까지 벌어지고 고등교육에서 효과가 다른 학교급보다 큰지 작은지도 메타분석마다 일치하지 않는다(Ⅱ장 4절 가). 비교 조건의 강도와 챗봇 세대가 섞인 평균이어서 대학의 판단 근거로는 비교 조건과 측정 방식이 도입 강좌와 비슷한 개별 실험(Ⅱ장 3절 나·다)이 더 적절하다.

나. 대학 강좌의 무작위 실험

하버드 대형 학부 물리 강좌(N = 194)의 교차설계 RCT는 GPT-4 기반 튜터로 집에서 학습한 조건을 같은 내용의 교실 능동학습이라는 강한 비교 조건과 견주었다. 수업 직후 검사의 선형회귀 효과크기는 0.63이었으나 저자들은 천장효과로 과소추정된 값이라 보았고 분위회귀로 추정한 효과크기는 0.73~1.3 SD였다. 적용 맥락은 블룸 분류의 이해·적용·분석 수준 학습 목표와 학생이 주제를 처음 본격적으로 배우는 초회 학습 상황이었다(Kestin et al., 2025). 튜터 개발팀이 수행한 연구이고 독립 재현은 연구진이 확인한 범위에서 찾지 못했다.

독일 뮌헨공과대학교(TUM)의 3집단 RCT(개발팀 수행)는 입문 프로그래밍(CS1) 강좌 학생 275명을 대상으로 90분 실습에서 보정된 힌트를 주되 완전한 해답은 보류하는 스캐폴딩 튜터(Iris), 무제한 ChatGPT, AI 없는 조건을 비교했다. 두 AI 집단은 통제집단보다 실습 점수가 크게 높았지만 사전·사후 지식 향상이나 코드 이해에서는 우위가 없었다. 내재 동기를 높인 조건은 Iris뿐이었으나 학생들은

ChatGPT 를 더 쉽고 도움이 된다고 평가했고 저자들은 이를 선호와 교수적 효과가 어긋나는 '편안함의 함정'이라 불렀다(Bassner et al., 2026).

국내에서는 서울 소재 대학 영어 강좌 학부생 80 명을 세 집단에 무작위 배정한 실험이 있다. 통제집단은 AI 없이 구두 요약 과제를 했고 두 처치집단은 각각 ChatGPT-3.5 에 튜터 역할을 맡기는 프롬프트와 요약 안내까지 포함한 더 구체적인 프롬프트를 썼다. 두 처치집단은 전반적 말하기 과제 수행에서 통제집단보다 유의하게 나았지만 전달력과 언어 사용 등 하위 요소에서는 유의한 효과가 없었다. 개입은 50 분 세 차례로 짧았고 연구팀은 효과가 프롬프트 설계에 따라 달라질 수 있다고 보았다(Sok & Shin, 2025).

큰 효과가 보고된 하버드 실험(개발팀 수행)은 강좌 내용에 맞춰 구조화한 튜터를 썼고 초회 학습에 한정됐다. 독일 실험에서는 실습 점수 상승이 지식 향상의 우위로 이어지지 않았고 학생이 선호한 조건과 동기를 높인 조건이 달랐다. 국내 실험의 효과는 전반적 말하기에 그쳤고 하위 요소에서는 나타나지 않았다. 비교 조건이 능동학습·AI 없음으로 다르고 개입 기간도 달라 효과크기를 직접 견주기 어렵다. 효과는 사용 중 수행이 아니라 AI 없이 치르는 지식 평가와 지연 측정으로 확인해야 한다.

이번 호에서 검토한 대학 실험 가운데 전용 튜터와 범용 AI 를 같은 강좌에서 직접 견준 연구는 독일 입문 프로그래밍 실험(개발팀 수행) 한 편이다. 이 실험에서 스캐폴딩 튜터와 무제한 ChatGPT 는 실습 점수를 함께 높였으나 두 조건 모두 지식 향상의 우위가 없었고 내재 동기는 스캐폴딩 튜터에서만 높았다. K-12 실험에서는 교사 설계 힌트를 준 GPT Tutor 에서 GPT Base 의 시험 성적 저하가 나타나지 않았다(Ⅱ 장 3 절 다). 이 실험들의 비교 조건은 범용 LLM 의 학습 모드가 아닌 범용 채팅이었다.

다. 대규모 현장 실험과 K-12 근거

실제 학기 운영 속 대규모 현장 실험의 결과는 달랐다. 미국 주립대 마케팅원론 학부생 약 500 명(분석 454 명)을 대상으로 한 무작위 현장 실험은 대면·비동기 온라인 두 방식에서 진행됐고 교수자 자료 기반 생성형 AI 챗봇(1 저자가 교수자이자 챗봇 콘텐츠 제작자)은 측정한 어떤 성과에도 통계적으로 유의한 영향을 주지 않았다. 비교 조건은 같은 참여 점수를 준 Quizlet 학습이었으며 시험 성취의 효과크기는 d = 0.015 였다(Thoeni & Fryer, 2026).

146 개국 5,831 명이 수강한 무료 온라인 코딩 강좌의 대규모 RCT(대학 정규 강좌 아님, 개발팀 수행)는 학생이 직접 풀이를 얻지 못하도록 연구팀이 프롬프트를 설계한 GPT-4 채팅 접근을 일부 학생에게 무작위로 제공했다. 채택자의 시험 점수 향상 추정치(효과크기 0.40)는 다중검정 보정 후 통계적으로 유의하지 않았다(Nie et al., 2025).

대학 일반화에 한계가 있는 K-12 연구에서도 사용 중의 향상과 이후 시험 성적은 따로 움직였다. 튀르키예 한 고등학교의 RCT 는 9~11 학년 약 50 개 학급(1,000 명 가까운 학생)에서 90 분 수업 4 회로 진행됐다. 문제 풀이 중 GPT-4 를 쓰게 하자 성적이 GPT Base 에서 48%, GPT Tutor 에서 127% 향상되었다. AI 없이 치른 시험에서는 가드레일 없는 GPT Base 집단이 통제집단보다 17% 낮았고 교사 설계 힌트를 주도록 지시한 GPT Tutor 집단은 통제집단과 통계적으로 구별되지 않았다. 학생들은 GPT Base 를 풀이를 베끼는 '목발'로 자주 썼다(Bastani et al., 2025).

나이지리아 고교 1 학년 영어 학습에 GPT-4 기반 Microsoft Copilot 을 6 주간 활용한 RCT(동료심사 전, K-12)는 종합 평가에서 0.31 SD, 주 성과인 영어에서 0.23 SD 의 향상을 보고했다. 세계은행 소속 저자들의 사업 평가이고 비교 조건은 무처치였다(De Simone et al., 2025). Khan Academy 의 Khanmigo 군집 무작위 실험(동료심사 전, K-12, 제공사 데이터 협력)은 미국 테네시주 18 개

중학교에서 2 년간 진행됐고 배정 효과는 학년도당 약 0.06 SD, 2 년차 약 0.08 SD 였다. 이 향상은 AI 지원 없이 같은 플랫폼을 쓸 때 예측되는 수준과 비슷했다(Oreopoulos & Low, 2026).

학기 운영 속 대학 강좌의 대규모 현장 실험(마케팅원론)에서 교수자 자료 기반 챗봇의 효과는 유의하지 않았다. 대학 정규 강좌가 아닌 무료 온라인 코딩 강좌의 대규모 RCT 에서도 채택자 효과는 다중검정 보정 후 유의하지 않았다. K-12 실험에서는 무처치와 비교한 나이지리아 실험을 빼면 효과가 작거나 시험 성적으로 이어지지 않았다. K-12 결과를 대학에 그대로 적용할 수는 없지만 대학 도입의 성과 역시 튜터 사용 중의 점수보다 AI 없이 치른 평가와 학습 참여 지표로 판단해야 한다(사용률은 Ⅱ장 4 절 라).

라. 교수법 설계 요소별 근거

Q1 의 질문 설계에 관한 근거는 튜터가 묻는 쪽보다 학생이 묻는 쪽에서 나왔다. 단일 회기 실험(실험실 과제, 배정 방식·효과크기는 초록에 없고 대화 상대와 질문 방향이 함께 다름)에서 한국 대학생 114 명은 생물 지문을 읽고 연습문항 2 개를 푼 뒤 동료·GPT·ITS 중 한 상대와 대화했다. 질문할 수 있었던 동료·GPT 집단은 시스템 질문에 답하기만 한 ITS 집단보다 대화 후 절대 점수와 조건부 정답률이 유의하게 높았다(Yang et al., 2025).

오개념을 겨냥한 설계의 근거는 서로 어긋난다. 반박 텍스트(refutation text) 메타분석에서 44 개 독립 비교(n = 3,869)의 전체 효과크기는 g = 0.41 이었고 출판 편향을 보정하자 g = 0.28 로 줄었다. 비교 대부분이 대학생 등 중등 이후 표본이지만 AI 튜터가 아닌 텍스트 개입 연구다(Schroeder & Kucera, 2022). 대학생 75 명이 분수 연산 오개념을 다루는 규칙 기반 컴퓨터 튜터로 학습한 집단 간 실험에서는 오개념이 드러난 오류 예제를 더한 튜터(39 명)와 일반 문제 풀이 튜터(36 명) 모두 향상되었지만 향상 폭에 유의한 차이가 없었다. 빈도주의·베이즈 분석 모두 오류 예제 추가의 이점을 강하게 지지하지 않았다(Di Lonardo Burr et al., 2020).

사용 강도와 학습자 특성에 관한 근거는 한 실험의 집단별 결과에 그친다. 연구팀은 생성형 AI 영어 말하기 앱을 개발한 한국 기업과 협력해 4 주 무작위 현장 실험(한국인 대학·대학원생 248 명, 주 분석 181 명)을 진행했고 성과 지표는 어휘 다양도(TTR) 변화율이었다. 무제한 AI 집단은 인간 원어민 튜터 집단보다 5.90% 더 향상해 유의했지만 시간 제한 AI 집단의 2.52% 향상은 통계적으로 유의하지 않았다. 무제한 AI 집단에서는 숙련도가 평균 이하인 학습자의 TTR 이 9.53% 유의하게 향상됐다(Cheon et al., 2025).

국내 연구 가운데 용인대 교양 일본어 수강생 74 명의 단일 강좌·자기 보고 연구는 교수자 설계의 질을 강조했다. 수강생들은 OpenAI 의 맞춤형 챗봇 기능인 GPTs 로 교수자가 설계한 튜터를 썼고 저자들은 생성형 AI 의 교육적 효과가 사용 빈도만이 아니라 배치의 교수법 설계 품질에 달려 있다고 보았다. 범용 GPT 와 실험적으로 비교하지는 않았다(H. Lee & Kwon, 2026). 대화 방식에 관해서는 출판 편향을 보정한 Laun 과 Wolff(2025)가 텍스트 기반 챗봇의 효과크기가 음성 기반 상호작용보다 유의하게 컸다고 보고했다.

Q1 의 세 설계 요소는 근거 수준이 다르다. 학생 질문을 끌어내는 대화 구조에는 초기 근거가 있으나 튜터가 묻는 소크라테스식 질문 자체를 대학 강좌에서 검증한 무작위 실험은 공개 자료에서는 찾지 못했다. 시스템이 질문만 던진 조건이 낮았다는 단일 회기 결과는 튜터가 묻기만 하는 설계의 한계를 시사하지만 대화 상대와 질문 방향이 함께 달라 확정할 수 없다. 오류 예제를 더한 규칙 기반 튜터 실험 1 편(대학생 75 명)에서는 향상 폭에 유의한 차이가 없었고 반박 텍스트 메타분석(AI 튜터 아님)은 출판 편향 보정 뒤에도 정적 효과를 보였다. 스캐폴딩을 구조로 넣은 튜터의 개발팀 수행 RCT 두 편은

결과가 서로 달랐다. 하버드 물리 튜터(부분별 순차 안내, Ⅱ장 1절 라)는 교실 능동학습보다 큰 효과를 보고했으나 독일 입문 프로그래밍의 Iris는 내재 동기를 높이고도 지식 향상의 우위가 없었다(Ⅱ장 3절 나). 두 튜터 모두 여러 요소를 묶은 설계여서 스캐폴딩만의 효과는 가를 수 없다. 음성 대화가 텍스트보다 유리하다는 결과도 없었다. 두 AI 조건을 직접 견준 검정은 보고되지 않았고 숙련도별 결과도 무제한 AI 집단 안의 하위집단 결과여서 사용 강도와 학습자 특성이 효과를 조절하는지는 검증되지 않았다. 도입 대학은 후보 튜터의 질문 방식과 오개념 대응을 직접 시험하고 효과는 학내 측정으로 확인해야 한다.

마. 교수자·조교 통합과 국내 대학 사례

강좌 운영에 AI 튜터·조교를 넣은 연구는 학업 성취보다 인식과 운영 경험을 주로 보고한다. 그 예로 2023년 가을 대형 온라인 대학원 컴퓨터과학 강좌에서 이뤄진 연구(설문 분석 337명)는 조지아공대 Jill Watson 접근 여부를 무작위 배정했다. Jill Watson은 교수 실재감의 설계·조직 차원과 사회적 실재감에 작지만 유의한 정적 효과를 보였고 인지적 실재감과 촉진·직접 교수 차원에는 효과가 없었다. 개발팀이 수행한 연구이며 성과는 학업 성취가 아닌 학기 말 탐구공동체(CoI) 설문의 실재감 인식이었다(Lindgren et al., 2024). 보고된 두 효과의 유의확률은 단측 검정 값과 일치해 양측 기준으로는 유의하지 않다(연구진 계산).

미국 대학 입문 파이썬 강좌의 한 학기 준실험(50명)에서는 연구팀이 개발한 LLM 조교 CodeTutor 집단의 학기 말 최종 점수가 사전검사 대비 평균 12.50점 오르고 통제집단은 3.17점 내렸다. 그러나 사전검사는 22문항 100점 만점이고 최종 성적은 실습·과제·중간고사·기말시험의 가중 평균으로 보너스 점수가 주어지면 100점을 넘을 수 있어 두 점수의 차이를 향상 폭으로 곧바로 읽기는 어렵다. CodeTutor를 쓸 수 없는 기말시험에서도 실험집단 안의 향상은 유의했지만 집단 간 표준화 효과크기는 보고되지 않았다. 학기가 진행되며 CodeTutor 제안에 대한 동의는 줄고 인간 조교 지원 선호가 커졌다(Lyu et al., 2024).

국내 대학 사례에서도 사람의 역할은 남았다. Ⅰ장 1절 나의 KAIST 가상 조교 강좌에서는 수강생 477명 가운데 연구에 동의한 472명 중 약 50%가 VTA를 한 번 이상 썼고 대화 916건·질의응답 3,869건이 발생했다. 개발팀 평가의 사전 설문에서 학생의 58%가 질문이 어리석어 보일까 하는 걱정 등으로 인간 교수자에게 질문을 망설인 적이 있다고 답했다. 같은 평가의 집단 비교(인과 분석 아님)에서 질문을 망설인 적이 있는 학생은 VTA를 더 편하게 느끼고(인간 조교 대비 편안함 점수 사전 0.69 → 사후 0.76, 망설인 적 없는 학생 0.42 → 0.47) 평균 13.2회로 망설인 적 없는 학생(7.8회)보다 자주 썼다. 그러나 VTA에 대한 신뢰는 인간 교수진보다 낮게 유지됐다(Kweon et al., 2025). 2025년 봄 학기 국내 한 대학 파이썬 입문 강좌에 배치된 SCALA(Ⅱ장 1절 라)에서는 조교가 학생을 오도하거나 지나치게 긴 사전 생성 질문·답변을 걸러냈고 학기 간 강의평가에서 체감 난이도가 2.44에서 1.74로 낮아졌다. 개발팀의 비무작위 비교이며 학습 효과 지표도 아니다(J. Lee et al., 2026).

학생의 생성형 AI 활용을 사람이 지도하는 모델은 A대학 글쓰기 교과 사례에서 근거를 얻는다. 이 교과는 생성형 AI 활용과 전문 튜터 면담을 결합한 인간 참여형(HITL) 수업을 운영했고 연구팀은 3개 학기 자료(성찰일지 634건 등)를 분석했다. 튜터의 교정적·지시적 피드백은 AI 생성 텍스트를 학술 규범에 맞게 정제하는 토대가 됐고 촉진적·질문형 피드백은 AI 결과물의 비판적 재구성을 유도했다. 통제집단이 없는 단일집단 연구다(김혜림·정향윤, 2026).

Jill Watson·CodeTutor·KAIST·SCALA 사례와 글쓰기 수업 연구에서 AI 튜터는 교수자에게 묻기를 망설이는 학생의 질문 창구가 될 수 있지만 신뢰·교정 피드백과 응답 검수에서는 사람의 역할이 남는다

. 근거 대부분이 개발팀 수행·비무작위·인식 측정이어서 운영 모델의 학습 효과는 검증되지 않았다. Q3의 운영 모델로는 즉시 응답과 일반 내용 질문은 AI 가, 응답 검수·심화 지도는 교수자·조교가 맡는 분담을 검토할 만하다(IV 장 1 절 다, IV 장 2 절 나). Jill Watson 이 사회적 실재감에 작은 정적 효과를 보인 결과(개발팀 수행, 인식 측정, 단측 기준 유의)는 AI 가 관계의 일부를 보조할 가능성도 보여주지만 학업 성취로 이어지는지는 확인되지 않았다. 연구별 설계·대상·효과크기는 [표 II-3]에 정리했다.

[표 II-3] 효과 실증 연구 요약

연구	설계	대상	표본	효과크기·주요 결과	근거 수준
VanLehn(2011)	실험 종합 리뷰	학교급 혼합	—	인간 튜터링 d = 0.79, ITS 0.76	메타
Kulik 과 Fletcher(2016)	메타분석(ITS)	학교급 혼합	통제 평가 50 건	효과 중앙값 0.66 SD	메타
Ma 외(2014)	메타분석(ITS)	전 학교급	효과크기 107 개	대집단 수업 대비 g = 0.42, 개별 인간 튜터링과 차이 없음	메타
Steenbergen-Hu 와 Cooper(2014)	메타분석(ITS)	대학생	39 개 연구	g = 0.32~0.37, 인간 튜터링보다 낮음	메타
Wu 와 Yu(2024)	메타분석(챗봇)	학교급 혼합	24 편(초록 기준)	전체 ES = 0.964, 고등교육 1.079	메타 (ChatGPT 이전 챗봇 중심)
Laun 과 Wolff(2025)	메타분석(챗봇)	학교급 혼합	62 편(효과크기 135 개)	g = 0.76, 출판 편향 보정 시 0.381	메타(LLM 이전 중심)
Alemdağ(2025)	메타분석(챗봇)	학교급 혼합	28 편	g = 0.48, 비교 조건·기간이 조절	메타(LLM 이전 중심)
Wang 외(2025)	메타분석(언어 학습 챗봇)	학교급 혼합	28 편	g = 0.484	메타
Kestin 외(2025)	교차설계 RCT(능동학습 대비)	하버드 학부 물리	194 명	분위회귀 0.73~1.3 SD(선형회귀 0.63)	중(개발팀 수행)
Bassner 외(2026)	3 집단 RCT(AI 없음 대비)	독일 대학 입문 프로그래밍	275 명	실습 점수 상승, 지식 향상 우위 없음	상(개발팀 수행)
Sok 과 Shin(2025)	3 집단 무작위 배정(AI 없음 대비)	서울 소재 대학 영어	80 명	전반적 말하기 향상, 하위 요소 효과 없음	중(국내)
Thoeni 와 Fryer(2026)	분반 내 RCT(Quizlet 대비)	미국 주립대 마케팅원론	454 명	시험 성취 d = 0.015, 유의한 효과 없음	상(개발팀 수행)

연구	설계	대상	표본	효과크기·주요 결과	근거 수준
Nie 외(2025)	대규모 RCT	무료 온라인 코딩 강좌	5,831 명	채택자 0.40(보정 후 비유의)	상(정규 강좌 아님·개발팀 수행)
Lindgren 외 (2024)	RCT(비접근 대비)	미국 온라인 대학원 CS	337 명	교수·사회적 실재감 일부에 작은 효과(단측 기준 유의)	상(개발팀 수행·인식)
Lyu 외(2024)	집단 간 준실험	미국 대학 입문 파이썬	50 명	사전검사 대비 최종 점수 +12.50 점(통제 –3.17 점, 척도 다름)	하(개발팀 수행)
Cheon 외(2025)	현장 RCT(인간 튜터 대비)	한국인 대학·대학원생 영어	248 명(분석 181 명)	무제한 AI 어휘 다양도 +5.90%, 시간 제한 AI 비유의	중(국내·기업 협력)
Di Lonardo Burr 외(2020)	집단 간 무작위 실험	캐나다 대학 분수 연산	75 명	오류 예제 추가 이득 없음	중(규칙 기반 튜터)
Bastani 외(2025)	현장 RCT	튀르키예 고교 수학	약 1,000 명	GPT Base 시험 17% 낮음, GPT Tutor 차이 없음	상(K-12)
De Simone 외 (2025)	RCT(무처치 대비)	나이지리아 고교 영어	처치 657 명·통제 671 명 배정(분석 처치 422 명·통제 337 명)	종합 0.31 SD, 영어 0.23 SD	참고(동료심사 전·K-12)
Oreopoulos 와 Low(2026)	군집 RCT	미국 중학교 수학	18 개 중학교	학년도당 약 0.06 SD, AI 없는 같은 플랫폼 예측과 비슷	참고(동료심사 전·K-12·제공사 협력)

자료: 각 연구 원문(일부 초록)을 바탕으로 연구진 정리

주: 근거 수준은 상 = 동료심사 RCT·표본 250 명 이상, 중 = 동료심사 무작위 배정 실험·표본 250 명 미만, 하 = 준실험, 메타 = 메타분석·종합 리뷰, 참고 = 동료심사 전 자료다. 개발팀 수행은 평가한 튜터를 연구자가 직접 개발한 경우다. 표본 칸의 —는 인용 자료에 표본 수가 없다는 뜻이다. ITS 메타분석 4 편과 Di Lonardo Burr 외(2020)는 LLM 이전 튜터를, 챗봇 메타분석 가운데 Wu 와 Yu(2024)·Laun 과 Wolff(2025)·Alemdağ(2025)는 LLM 이전(ChatGPT 확산 이전) 챗봇을 주로 다뤘다.

소결 LLM 이전 ITS 의 대학 효과는 중간 수준(g = 0.32~0.37)이었다. LLM 튜터를 쓴 대학 강좌 실험의 결과는 비교 조건(능동학습·AI 없음·Quizlet 학습), 측정(직후 검사·강좌 성적·인식 설문), 설계(강좌 자료 기반 구조화·정답 보류·범용 채팅)에 따라 큰 정적 효과부터 무효과까지 흩어졌다. 큰 정적 효과는 강좌 내용에 맞춰 구조화한 하버드 물리 튜터 실험에서 보고됐고 이해·적용·분석 수준 학습 목표의 초회 학습이었다. 그러나 같은 단기 구조화 조건인 독일 입문 프로그래밍 90 분 실습에서는 지식 향상의 우위가 없어 인지 수준·기간에 따른 효과의 경계는 판단할 수 없다. 학기 운영 속 대학 현장 실험(마케팅원론)에서는 효과가 없었다. 큰 효과를 보고한 하버드 실험은 개발팀 수행 연구이며 K-12 결과는 대학에 그대로 적용하기 어렵다. 두 대표 솔루션은 제공사와 독립된 학습 효과 근거가 확인되지 않아(Ⅲ장 2 절 가·나) Ⅱ장 3 절의 근거는 개별 제품이 아닌 튜터 설계 유형을 대상으로 한다. 핵심 질문 Q2 의

잠정 답은 '좁은 조건에서 확인된 효과'에 가깝고 과목·인지 수준·기간·학습자 특성별 한계는 Ⅱ장 4 절 가에서 다룬다.

4. 한계·논쟁점

효과의 크기는 Ⅱ장 3 절에서, 국내 법제는 Ⅳ장 1 절에서 다루므로 Ⅱ장 4 절은 쟁점의 성격과 대응 수단을 중심으로 살피고 요지를 [표 Ⅱ-4]에 모은다.

가. 효과의 일반화 한계

일반화의 첫 경계는 과목과 인지 수준이다. 하버드 물리 RCT(Ⅱ장 3 절 나)의 저자들은 여러 개념의 복합적 종합이나 고차 비판적 사고가 필요한 맥락 등 모든 맥락에서 구조화된 AI 튜터링이 교실 능동학습보다 항상 우월하다고 가정하지 않는다고 밝혔다(Kestin et al., 2025). GPT-4 기반 튜터를 평가한 튀르키예 고교 RCT(K-12)도 한 고등학교의 수학 단일 주제에 한정되었고 단기 성과만 측정했다(Bastani et al., 2025).

기간에 따른 결과는 메타분석마다 달랐다. 포함 연구 대부분이 ChatGPT 확산 이전 챗봇인 한 메타분석(Ⅱ장 3 절 가)에서는 10 주 미만 단기 개입의 효과크기(ES = 1.179)가 10 주 이상 장기 개입(ES = 0.492)보다 컸다. 저자들은 이를 장기 개입에서 사라지는 챗봇의 신기성 효과로 설명했다(Wu & Yu, 2024). 출판 편향을 보정한 다른 챗봇 메타분석에서는 반대로 1 회성 실험보다 2 일~1 개월·3 개월 초과 개입의 효과크기가 유의하게 컸다. 같은 메타분석은 6 개월을 넘는 장기 연구가 없어 신기성 효과와 지속 효과를 충분히 평가하지 못했다고 밝혔다(Laun & Wolff, 2025). 대학 ITS 메타분석에서는 연구 시점에 따른 차이가 나타나 초기 연구의 효과가 최근 연구보다 유의하게 컸다(Steenbergen-Hu & Cooper, 2014).

학습자 특성에 따른 결과는 한 방향으로 모이지 않는다. Ⅱ장 3 절 가에서 본 챗봇 메타분석은 고등교육 학습자 집단의 효과크기는 크고 유의했으며 초·중등교육보다 컸다고 보고했다(Wu & Yu, 2024). 반면 출판 편향을 선택 모형으로 보정한 메타분석에서는 고등교육 환경의 효과크기가 초등교육보다 낮았다(Laun & Wolff, 2025). 숙련도가 평균 이하인 학습자의 향상도 한 실험의 하위집단 결과에 그친다(Ⅱ장 3 절 라).

효과의 크기는 측정 방식과 비교 조건에 따라 바뀐다. 여러 학교급을 아우른 ITS 메타분석에서 개선 폭은 자체 개발 시험과 표준화 시험 중 무엇으로 쟀는지에 크게 좌우되었으며 저자들은 시험과 수업 목표의 정렬을 핵심 결정 요인으로 보았다(Kulik & Fletcher, 2016). 챗봇 메타분석에서는 비교 집단에 별도 지원이 없고 실험이 1 회기뿐이며 챗봇이 과업 중심으로 자주 묻는 질문을 처리할 때 효과크기가 가장 컸다(Alemdağ, 2025).

보고된 수치 자체가 얼마나 안정적인지도 따져 본다. Ⅱ장 3 절 가에서 본 한 챗봇 메타분석에서는 출판 편향을 선택 모형으로 보정하자 전체 효과크기는 g = 0.76 에서 g = 0.381 로 줄었지만 유의했다(Laun & Wolff, 2025). 단기 개입의 효과가 더 컸던 챗봇 메타분석에서는 출판 편향이 확인됐고 소규모 연구 편향의 영향 가능성이 제기됐다(Wu & Yu, 2024).

과목·기간·학습자 특성·측정 방식별 근거에서 대화형 튜터의 긍정적 결과가 주로 보고된 대상은 구조화된 과목의 초회 학습과 짧은 개입이었으나 같은 입문 과목에서도 결과가 갈렸다. 고차 사고 과제, 한 학기를 넘는 사용, 학습자 특성에 대한 근거는 얇고 개입 기간의 효과와 고등교육의 상대적 효과도 메타분석마다 일치하지 않아 수치를 신중히 읽어야 한다. 핵심 질문 Q2 에 답하려면 학내 도입 효과를 고차 사고 과제, 장기 사용, 표준화된 측정으로 확인해야 한다(Ⅳ장 2 절 나).

나. 정답 노출과 우회 사용

정답을 곧바로 주지 않도록 만든 튜터도 우회 요청에는 약했다. 선형대수 벤치마크로 교육용 LLM 을 평가한 연구팀(Ⅱ장 1절 라)은 '교수법적 탈옥'을 중대한 취약점으로 지목했다. 예비 실험에서 암호(Cipher) 같은 '불일치 일반화' 공격은 차단됐지만 거부 억제나 역할극(예: '아픈 할머니') 같은 '경쟁 목표' 공격은 높은 성공률을 보였다(S. Zhao et al., 2026).

실제 학생 없이 적대적 학생 에이전트를 쓴 평가에서는 대부분의 기법이 모든 모델에서 높은 정답 노출을 일으켰고 대인 영향 기법은 교수법 정렬 튜터 TutorRL-7B 에서 최대 85%에 이르렀다. Ⅱ장 1절 나에서 본 확장 실험의 GPT-5 는 누출 응답이 4.58%였으나 정답을 숨겨도 지나치게 많은 정보를 주는 경향이 있었다(J. Zhao et al., 2026). 시뮬레이션 학생으로 훈련한 교수법 정렬 강화학습 실험에서는 학생 해결률 향상이 가장 큰 Qwen2.5-72B 와 DeepSeek 의 DeepSeek V3 가 정답 노출도 많았다(Dinucu-Jianu et al., 2025).

대학 현장의 경고는 같은 쪽을 가리킨다. 12 개국 43 개 대학의 생성형 AI 지침을 분석한 연구에서 코넬대 지침은 탈옥이 금지 주제 우회뿐 아니라 과제 정답을 얻는 데도 쉽게 쓰일 수 있다고 경고했다(Ng et al., 2026). CS50 개발팀은 잦은 코드 블록 생성(Ⅱ장 1절 나)이 학생의 CS50 Duck 의존을 키우고 문제 해결 과정 참여를 막을 수 있다고 우려했다(R. Liu et al., 2025).

국내 사례로는 학생 코드를 분석하고 안내하는 엘리스의 AI 헬피가 있다. 교육자 매뉴얼에 따르면 AI 헬피의 실습 화면 예시 기능 가운데 코드 실행 요청은 실습 자료의 정답 코드를 바탕으로 실행된다. 이 기능은 정답 코드에 도달하기 위한 사고 단계도 안내한다(엘리스, Retrieved 2026). 정답 코드가 학생에게 어디까지 드러나는지는 제공사 문서만으로 알 수 없어 실측을 거쳐야 확인된다.

정답 보류는 벤치마크·시뮬레이션 평가와 대학 현장의 경고가 보여 주듯 적대적 요청, 프로그래밍 같은 절차적 과목, 기반 모델 교체(Ⅱ장 1절 나)에서 흔들리기 쉽고 학생 해결률을 크게 올린 모델이 정답도 많이 노출하는 상충도 있다. 튜터 설정에만 기대서는 학업 정직성을 지키는 데 한계가 있으므로 강의실 대면 퀴즈 같은 평가 설계(Ⅲ장 1절)와 응답 사후 점검을 병행하는 편이 현실적이다.

다. 환각과 교과 내용 오류

논의는 학습자가 AI 응답의 오류를 가려낼 수 있는지에서 시작한다. 한 연구에서는 독일 대학 물리학 1·2 학년생 102 명이 입문 역학 3 개 문항에 대한 ChatGPT 응답(2023 년 GPT-3.5)과 모범답안을 평가했는데 제시된 ChatGPT 응답은 모두 틀리거나 부정확·불완전·오도적이었다. 그런데도 대부분 학생이 모르는 최고 난도 문항에서는 ChatGPT 응답이 모범답안만큼 좋게 평가됐고 자기평가한 전문성이 낮을수록 두 응답의 정확성 평가 차이가 작았다. 연구팀은 응답이 잘 쓰였다는 인식이 과학적 정확성 판단을 덮을 수 있다고 보았다(Dahlkemper et al., 2023).

강좌 전용 튜터에서도 오답은 나왔다. CS50 개발팀 논문에 따르면 개발팀과 별개인 선임 강좌 직원이 2023 년 여름 학기 CS50 Duck 의 Ed 게시판 답변을 검토한 결과 교과 답변은 25 건 중 22 건(88%), 행정 답변은 39 건 중 30 건(77%)이 정확했다. 2023 년 가을 학기 Ed 답변 180 건 가운데 직원 승인을 받은 답변은 70 건(39%)이었으나 저자들은 학생의 Ed 이용이 크게 줄어 이 계산이 실제 정확도를 과소평가한다고 보았다(R. Liu et al., 2024). 교육 분야 RAG 서베이는 남은 과제로 환각 완화, 검색 지식의 완전성·시의성 확보, 계산 비용 절감, 멀티모달 지원 강화를 꼽았다(Li et al., 2025). 조지아공대 Jill Watson 개발팀은 제한된 문맥으로 답해야 하는 RAG 의 한계와 안전을 위해 답할 수 있는 질문까지 거르는 성능 상충을 한계로 밝혔다(Taneja et al., 2024).

한 모델 안에서 교과 정확성과 교수법이 충돌할 때가 있다. 고등학교 수학 다단계 문제로 만든 튜터 벤치마크(K-12 내용)에서는 교과 전문성과 교수법 능력 사이에 튜터 특화 정도에 따른 상충 관계가 나타났다. 교수법에 특화하면 문제 풀이 능력이 떨어지고 풀이 정확도가 높은 모델은 교수법이 부족한 경우가 많았다(Macina et al., 2025). 대학 명제논리 튜터의 증명 상태로 피드백 에이전트를 평가한 연구(II 장 1 절 다)에서는 정확한 진단이 교수법적으로 실행 가능한 피드백으로 이어지지는 않았다(Yasir et al., 2026). 두 연구 모두 실제 학생의 학습을 잰 것은 아니다.

음성 튜터의 발음 피드백에서는 음성 인식의 과잉 교정이 문제로 지적된다. Speak 소속 저자의 기술 평가(II 장 1 절 마)에서 음성 인식 8 종의 과잉 교정률은 단어오류율과 음의 상관을 보여 단어오류율이 낮은 시스템일수록 발음 오류를 더 많이 가렸다. NVIDIA 의 음성 인식 모델 Parakeet 으로 speechocean762 말뭉치를 분석한 결과 발화의 82.0%에서 상위 가설에 과잉 교정된 발음 오류가 하나 이상 있었고 시스템은 과잉 교정할 때 가장 확신이 높은 경향이 있었다(Park, 2026).

그럴듯한 오류는 학습자가 스스로 판단하기 힘든 고난도 내용에서 더 위험해지는 것으로 보인다. 교과 정확성과 교수법 사이의 상충은 모델을 고르는 단계부터 점검할 문제다. 튜터 응답에 근거 자료를 표시하고 오류 신고 경로를 두며 고난도 내용에는 사람의 검토를 병행하는 방식이 대응 수단이 될 수 있다(평가척도 B2 콘텐츠 품질·정확성과 C1).

라. 낮은 자발적 사용

튜터를 제공하는 것과 학생이 실제로 쓰는 것은 다르다. K-12 에서는 Khan Academy 가 Khanmigo 에 접근할 수 있는 학생 가운데 약 15%만 이용한다고 밝혔다(제공사 발표, 이용의 정의와 측정 기간은 밝히지 않음). 제공사는 Khanmigo 초기 사용 양상이 고르지 않았고 대화마다 학생을 진전시키는 정도가 달랐다고 덧붙였다(Khan Academy, 2026). 같은 제품을 미국 중학교 보충수학에 쓴 군집 무작위 실험(동료심사 전, K-12, 제공사 데이터 협력)은 정의가 다른 사용 지표를 보고했다. 이 실험에서는 학생의 96%가 Khanmigo 에 한 번 이상 메시지를 보냈지만 중앙값 학생은 실수한 연습 세션의 17%에서만 메시지를 보냈다. 저자들은 효과가 주로 구조화된 개별 연습에서 나왔고 튜터 기여는 낮은 사용 강도로 제한되며 사용은 설계에 반응한다고 보았다(Oreopoulos & Low, 2026).

대학과 성인 온라인 강좌의 실험에서도 사용률은 낮았다. 학점 과정이 아닌 무료 자율 수강 온라인 코딩 강좌의 RCT(II 장 3 절 다, 개발팀 수행)에서 GPT-4 채팅 접근권을 받은 학생 가운데 실제로 사용한 학생은 14.2%에 그쳤다. 실험집단의 진단 시험 응시율은 44.1%로 통제집단 48.5%보다 낮았고 원문이 추정한 차이는 4.3%p 였다(Nie et al., 2025). 미국 한 주립대 마케팅원론 강좌의 무작위 현장 실험(II 장 3 절 다)에서 학생의 챗봇 로그인은 주당 평균 0.89 회로 과제로 요구한 주 1 회에 약간 못 미쳤다. 저자들은 학생이 AI 챗봇에 만족한다고 해서 AI 챗봇으로 학습한 것은 아니라고 지적했다(Thoeni & Fryer, 2026).

선택 자원으로 둔 도우미는 쓰지 않았다는 응답이 다수였다. 미국 한 온라인 대학의 2024 년 여름 1 학기 강의평가 883 건 가운데 약 68%가 AI 강좌 도우미 Spark 를 쓰지 않았다고 답했고 미사용 이유로는 '필요성을 느끼지 못함'이 252 건으로 가장 많았다. Spark 는 사용 의무가 없는 선택 자원으로 도입돼 강좌 활동의 핵심 요소가 아니었고 저자들은 실질적 거부보다 효과적이지 못한 소통(안내)이 핵심 문제라고 해석했다. 도입 대학 소속 저자의 내부 평가이며 응답이 고유 학생 수와 같지 않아 중복 응답이 섞일 수 있다(Hanshaw & Sullivan, 2025).

국내 사례에서는 참여 유지가 과제였다. A 대학교가 AI 기반 말하기 앱 플랭(PLANG)으로 3 개월 운영한 비교과 프로그램을 분석한 단일집단 연구에서 평균 이수율은 55.1%였고 운영진은 수료율 공지·학습 목표 이벤트·시상품으로 자발적 참여를 독려했다(황영아·신종호, 2025).

선택형으로 도입한 튜터는 실제 사용자가 일부에 그쳐 배정 집단 전체로 잰 효과가 희석될 가능성이 크다. 사용을 늘리는 설계의 효과는 아직 검증이 부족하지만 수업 활동에 튜터 사용을 결합하고 사용 목적을 안내하는 일이 핵심 질문 Q3 의 운영 모델에서 중요한 변수가 될 수 있다(IV 장 2 절 나).

마. 정서적 의존과 과의존

정서적 의존에 관한 근거는 주로 상관 연구에서 나왔다. Healthy Minds Study 2024-2025 자료로 미국 대학생 858 명의 AI 동반자 사용 예측 요인을 분석한 연구에서 응답 대학생의 17%가 동반자 목적으로 AI 를 쓴다고 답했다. 외로움은 전반적 AI 사용이 아니라 동반자 목적 AI 사용과만 독립적으로 관련됐으며 그 오즈비(OR)는 1.21 이었다. 아시아계·유학생은 AI 를 동반자로 쓸 가능성이 더 높았다. 다만 횡단 설계라 인과 추론이 불가능하고 측정 문항은 동반자 전용 앱과 범용 AI 도구를 구분하지 않았다(C. H. Liu et al., 2026).

문헌 119 편을 대상으로 한 범위 검토(스코핑 리뷰)는 교육에 한정하지 않고 LLM 챗봇의 정신건강 위해를 다뤘는데 여기서도 정서적·사회적 의존과 철회 증상을 보이는 문제적 챗봇 사용은 정신건강 증상과 상관이 있었다. 이 범위 검토가 인용한 체계적 문헌고찰(37 편)은 의인화, 정서적 투사, 개인화·적응 같은 설계 특성을 유사 사회적 유대 형성 요인으로 꼽고 이 유대가 정서적 의존과 사회적 고립 위험으로 이어진다고 보았다(재인용). 범위 검토의 저자들은 AI 의존 연구가 질적 연구와 양적 상관 연구 중심이며 적절한 실험 연구와 검증된 측정 도구가 부족하다고 정리했다(Diel et al., 2026).

인지적 과의존에는 실험 근거가 일부 있다. 10 개 강좌 학생 1,625 명을 대상으로 동료 피드백 맥락의 AI 지원이 학생 주체성에 미치는 영향을 본 무작위 통제 실험에서 학생들은 AI 지원에서 배우기보다 AI 지원에 의존하는 경향을 보였다. AI 지원을 없애자 같은 품질의 피드백을 쓰는 데 어려움을 겪었고 자기점검 전략으로 대체하면 공백을 일부 메웠지만 AI 지원만큼 효과적이지 않았다. 이 실험은 대화형 튜터가 아닌 피드백 작성 지원을 다뤘고 연구용 플랫폼의 개발자가 수행했으며 학기 초 8 주에 한정됐다(Darvishi et al., 2024).

OECD 는 정답을 바로 주는 생성형 AI 에 과의존하면 능동적 참여가 줄어 과제 수행은 나아져도 학습 성과는 따르지 않을 수 있고 과의존이 학생을 수동적 소비자로, 교사를 감독자로 만들 위험이 있다고 보았다(OECD, 2026a). CS50 Duck 을 쓴 자교 강좌를 혼합연구로 살핀 뮌헨응용과학대 연구팀(제공사와 독립)도 숙련 학생은 생성형 AI 도구를 문제 해결 강화에 쓰지만 어려움을 겪는 학생은 과의존해 핵심 학습 과정을 건너뛸 수 있다고 보고했다(Zönnchen et al., 2025).

의인화와 개인화를 강조하는 '동반자'형 설계는 정서적 의존 위험을, 정답을 바로 받는 사용 방식은 인지적 과의존 위험을 키울 소지가 있다. 정서적 의존 쪽은 상관 근거에 머물러 인과를 단정할 수 없으므로 튜터의 의인화 수준과 정서적 기능을 설계 단계에서 점검하고 자기점검 전략을 병행하는 방안은 검토할 가치가 있다.

바. 한국어 성능 격차

한국어 성능에 관한 근거는 대부분 튜터가 아닌 LLM 평가에서 나온다. NAVER Cloud 소속 저자들이 만든 KoNET 은 초졸·중졸·고졸 검정고시와 대학수학능력시험 4 종으로 구성된 한국 교육시험 기반 벤치마크다(K-12 시험, 자사 모델 포함 평가). 저자들에 따르면 기존 벤치마크는 영어 중심이어서

한국어 같은 저자원 언어의 성능을 잘 보여 주지 못한다. 2024 년 중반 모델로 평가한 결과 폐쇄형 상용 API 와 오픈소스 모델 사이의 성능 격차가 컸고 모델 성능은 학교급이 올라갈수록 떨어졌다. 『용비어천가』 기반 문항에서는 문화적 맥락 이해가 부족한 모델이 문항 해석에 실패했다(Park & Kim, 2025).

한국어 의료윤리 객관식 문항과 그 영어 번역본으로 ChatGPT 3.5·4.0·4.5 를 비교한 연구(학습자 대상 연구 아님)에서 버전 4.5 는 두 언어 모두에서 거의 만점에 가까웠던 반면 버전 3.5 는 특히 한국어 시험에서 성능이 제한적이었다. 언어 간 성능 격차는 모델 업그레이드로 줄었지만 4.5 에서도 일부 문항에서 통계적으로 유의하게 남았고 저자들은 영어에서 얻은 성능 향상이 한국어 같은 중간 자원 언어로 자동으로 이어지지 않는다고 해석했다(Kim et al., 2025).

2026 년 기준 한국어 교정 도구 5 종(네이버·다음·바른한글·Gemini 3·ChatGPT 5.2)을 비교한 연구(튜터 아님)에서는 LLM 기반 도구가 90 개 문항의 전체 정답률에서 앞섰다. Google 의 Gemini 3 은 92.2%, ChatGPT 5.2 는 71.1%로 규칙·통계 기반 검사기(46.7~53.3%)보다 높았다. 표기 차원에서는 ChatGPT 5.2 의 교정률이 73.33%로 Gemini 3(93.33%)과 네이버 검사기(80.0%)보다 낮았다. LLM 기반 교정 도구에서는 틀린 근거를 유창하게 정당화하는 문법적 환각과 사후 합리화가 확인됐고 가장 성능이 높은 Gemini 3 도 틀린 답을 확고한 규범적 사실처럼 제시하는 권위적 환각을 보였다(김영일, 2026).

국내 강좌 사례로는 국내 한 대학의 파이썬 입문 강좌에 배치된 SCALA 가 있다. 강좌 자료와 안내는 모두 한국어로 제공됐고 개발팀은 배치용으로 Mistral AI 의 24B 오픈소스 모델 Mistral 3 Small 을 골랐다. 이 모델은 이전 학기 학생 질문으로 시험했을 때 응답 중간에 한국어와 영어를 오가는 등 언어 사용이 일관되지 않고 지나치게 장황했다(J. Lee et al., 2026).

제공 범위의 차이도 있다. 제공사 문서인 Duolingo Max 도움말이 밝힌 Explain My Answer·Roleplay 제공 과정은 영어 사용자용 프랑스어·스페인어·독일어·포르투갈어·이탈리아어 과정과 스페인어·포르투갈어·일본어·독일어·프랑스어 사용자용 영어 과정이다. 한국어 사용자용 영어 과정은 이 목록에 없으며 문서 간 차이는 Ⅲ장 1 절에서 다룬다(Duolingo, Retrieved 2026-a).

이 연구들은 튜터 자체가 아닌 LLM 의 시험 성적이나 규범 설명을 잰 경우가 많아 튜터 품질로 곧바로 옮길 수는 없다. 그래도 한국어 콘텐츠와 한국인 학습자 맥락의 성능을 영어 성능으로 짐작할 근거는 부족하다. 틀린 답을 유창하게 정당화하는 오류는 정답률이 높은 모델에서도 나타나 학습자에게 더 위험할 수 있으므로 국내 도입 전에는 한국어 강좌 자료와 한국어 질문으로 먼저 실측해 보는 것이 순서다(Ⅳ 장 1 절 다, Ⅳ 장 2 절 나).

사. 데이터·규제와 근거의 질

대화형 튜터는 학생의 대화 기록을 다루므로 데이터 보호와 연령 기준이 쟁점이 된다. UNESCO 는 2023 년 교육·연구 분야 생성형 AI 지침에서 대부분 국가에 생성형 AI 규제가 없어 이용자 데이터 프라이버시가 보호되지 않고 교육기관도 도구 검증 준비가 부족하다고 진단했다. 또 생성형 AI 이용은 거의 언제나 이용자 데이터를 제공사와 공유하는 일을 수반한다고 보았다. 같은 지침은 생성형 AI 플랫폼과의 독립적 대화에 최소 13 세의 연령 기준을 권고했고 1:1 언어 코치형 활용에서도 문화적으로 부적절하거나 연령에 맞지 않는 출력 때문에 연령 제한을 둘 수 있다고 했다. 2023 년 문서여서 이후의 규제 변화는 반영되지 않았다(Miao & Holmes, 2023).

12 개국 43 개 대학 지침 분석(Ⅱ 장 4 절 나)에서는 지침이 다루는 범위가 고르지 않았다. 개인·민감정보 문제는 대학 지침 30 곳이 다뤘지만 악의적 이용 가능성을 언급한 곳은 8 곳, 탈옥을 포함한 취약점·

버그·익스플로잇을 다룬 곳은 4곳뿐이었다. 대학 지침은 프라이버시·보안 책임을 학생·교직원 개인에게 넘기는 경향을 보였다(Ng et al., 2026).

음성·영상 기능을 갖춘 튜터라면 감정 인식 규제와 닿을 수 있다. EU 집행위원회의 금지 관행 지침은 비구속적이며 EU 인공지능법(AI Act)의 최종 해석 권한은 EU 사법재판소에 있다. 이 지침은 교육기관이 학생의 흥미·주의를 추론하려고 감정 인식 AI를 쓰는 것을 금지 대상으로 본다. 교육기관 밖에서 쓰면 금지 대상이 아닌 감정 인식 온라인 언어학습 앱도 교육기관이 학생에게 사용을 요구하면 금지되며 글로 쓴 텍스트에서 감정을 추론하는 AI는 생체정보 기반이 아니어서 금지 범위 밖이다(European Commission, 2025).

제공사 공시도 생성형 AI의 위험을 다룬다. Duolingo의 2025 회계연도 연차보고서는 출력 정확성 문제·의도치 않은 편향·지식재산과 프라이버시 침해 가능성을 위험 요인으로 적었다(Duolingo, 2026b).

근거의 질에서는 철회 사례가 나왔다. Health Education Journal은 출판사(Sage)와 편집인 요청으로 간호학과 1학년 대상 방탈출 챗봇 학습 효과 논문(2025)을 2026년 8월 철회했다. 철회는 마지막 저자 소속의 진위에 대한 독자의 문제 제기에서 시작됐고 철회 사유는 마지막 저자의 신원을 확인할 수 없었던 점이며 저자들은 철회 결정에 동의하지 않았다. 철회된 원논문은 간호학과 1학년 48명(각 24명) 표본에서 Cohen's d = 1.52의 큰 효과를 보고했었다("Retraction: Gamified Mobile Learning," 2026). 이 값은 효과 근거로 세지 않는다.

대화 기록과 음성 데이터를 다루는 튜터는 프라이버시·연령·감정 인식 규제와 맞물리고 대학 지침이 기술적 위험을 개인 책임으로 넘기면 관리 공백이 생길 수 있다. 학습자 집단별로 튜터 응답 품질이 다른지를 직접 잰 고등교육 연구는 이번 호 자료에서 찾지 못했다. 제공사 공시가 의도치 않은 편향을 위험 요인으로 적었고 언어별 성능 차이가 불평등한 혜택을 낳을 수 있다는 지적(IV장 1절 다)도 있어 형평성 점검이 필요하다. 철회 사유는 저자 신원 문제였지만 소표본 연구의 큰 효과크기는 눈에 띄기 쉬운 만큼 정정·철회 여부와 독립 재현을 확인한 뒤에 받아들여야 한다. 국내 법제(고지·위탁·국외 이전)는 IV장 1절 가·나에서, 계약 점검은 부록 4에서 다루고 쟁점별 요지는 [표 II-4]에 정리했다.

[표 II-4] 주요 쟁점 정리

쟁점	확인된 사실(요지)	불확실성	대응 수단
효과의 일반화	정적 결과는 초회 학습의 구조화 튜터 실험에서 주로 보고, 개입 기간의 효과는 메타분석마다 방향이 엇갈림(Wu와 Yu(2024)는 10주 미만 우세·ChatGPT 이전 챗봇 중심, Laun과 Wolff(2025)는 1회성보다 긴 개입 우세)	고등교육의 상대적 효과가 메타분석마다 상충, 한 메타분석의 포함 연구에 6개월 초과 연구 없음(대학 장기 근거 부족)	고차 사고 과제·지연 측정·비교 집단을 둔 학내 효과 측정
정답 노출·우회 사용	'경쟁 목표' 공격의 높은 성공률(벤치마크), 대인 영향 기법의 정답 노출 최대 85%(TutorRL-7B, 시뮬레이션), 확장 실험 GPT-5 누출 응답 4.58%, 대학 지침의 탈옥 경고, 지시 희석에 따른 코드 블록 응답 증가(개발팀 수행, 현재 수치 미확인)	실제 학생 사용에서의 정답 노출 규모	응답 사후 점검, 정답 노출 실측, 대면 평가 병행

쟁점	확인된 사실(요지)	불확실성	대응 수단
환각·교과 내용 오류	틀린 응답을 고난도 문항에서 모범답안만큼 평가(GPT-3.5, 2023), CS50 Duck 게시판 답변의 직원 검토 정확도 교과 88%·행정 77%(2023 년 여름, 개발팀 수행), 교과 전문성과 교수법의 상충(벤치마크), 발음 오류 과잉 교정	최신 모델의 교과별 오류율	강좌 자료 근거·출처 표시, 오류 신고 경로, 사람 검토
낮은 자발적 사용	접근 학생 가운데 약 15% 이용(제공사 발표·K-12, 이용 정의 미공개), 실험의 실제 사용 14.2% (무료 온라인 강좌), 선택형 도우미 미사용 응답 약 68%	사용을 늘리는 설계의 효과	수업 활동과 결합, 사용 목적 안내
정서적 의존·과의존	대학생 17%가 동반자 목적 AI 사용(횡단 상관), AI 지원 제거 뒤 피드백 작성 어려움(무작위 실험, 튜터 아님)	정서적 의존은 상관 연구 중심이라 인과 불명	의인화 수준 점검, 자기점검 전략 병행
한국어 성능 격차	한국 교육시험의 모델 간 격차·학교급별 하락(K-12 시험), 의료윤리 문항의 언어 간 격차 일부 잔존, 문법적·권위적 환각, 국내 강좌 튜터 모델의 한국어·영어 혼용(개발팀 시험)	한국어 강좌 내용에 대한 튜터 정확도 평가 부족	한국어 강좌 자료·질문으로 실측
편향·형평성	제공사 공시의 의도치 않은 편향 위험 요인(Duolingo), 언어별 성능 차이에 따른 불평등 지적	학습자 집단별 튜터 응답 품질을 직접 잰 고등교육 연구 미확인	집단별 사용률·성과 점검
데이터·규제	43 개 대학 지침 가운데 취약점 언급 4 곳·악의적 이용 언급 8 곳, 책임의 개인 전가 경향, 교육기관의 감정 인식 AI 금지(EU 비구속 지침)	국내 대화 기록 보관 기준	계약상 보관·2 차 이용 확인, 연령·고지 점검
근거의 질	챗봇 메타분석의 출판 편향(보정 시 g = 0.76 → 0.381), 고등교육 챗봇 연구 철회(2026. 8.)	소표본 큰 효과크기의 재현성	정정·철회 확인, 독립 재현 확인

자료: Ⅱ장 4 절에서 인용한 연구·정책·제공사 자료를 바탕으로 연구진 정리

주: K-12·동료심사 전·개발팀 수행·제공사 발표·벤치마크·시뮬레이션은 근거의 성격 표시이며 불확실성은 연구진 판단이다.

소결 Ⅱ장 4 절의 쟁점([표 Ⅱ-4])은 효과의 일반화 범위, 정답 노출과 우회 사용, 환각과 교과 정확성, 낮은 자발적 사용, 정서적·인지적 의존, 한국어 성능 격차, 편향·형평성, 데이터·규제와 근거의 질로 정리된다. 무작위 실험에서 직접 관찰된 쟁점은 낮은 사용률과 AI 지원을 없앤 뒤의 피드백 작성 저하 등이고 정답 노출과 교과 오류는 주로 벤치마크·시뮬레이션에, 정서적 의존은 상관 연구에 기대므로 위험의 크기는 근거의 성격에 비추어 가늠해야 한다. 쟁점 다수는 튜터 기능을 고치는 것만으로 해소되기 어렵다고 연구진은 판단한다. 이 쟁점들은 Q2 의 일반화 범위를 좁히고 Q3 의 운영 모델에서 수업·평가 설계와 사람의 검토를 전제 조건으로 만든다. 국내 법제와 운영 모델은 Ⅳ장 1 절에서 이어서 다룬다.

Ⅲ. 대표 솔루션 심층 분석

Ⅲ장은 Ⅱ장의 솔루션군 분석에 이어 Ⅰ장 3절에서 고른 하버드 CS50 강좌의 CS50 Duck과 Duolingo의 Duolingo Max를 프로필, 평가척도 적용 결과, 비교 종합의 순서로 살펴본다. 판정은 보증이나 배척이 아닌 도입 검토 의견이다. 이번 호의 판정은 자료 기준일까지 공개된 문서와 연구에 근거하며 실측이 필요한 항목은 판정하지 않았다(Ⅰ장 4절). 점수도 판정 범위가 다른 진단 정보이며 순위를 나타내지 않는다.

1. 솔루션 프로필

CS50 Duck은 하버드 CS50 강좌 팀이 만든 강좌 전용 튜터이며 이용 범위는 강좌 정책으로 정해진다. CS50 강좌들은 과목이 명시적으로 허용한 경우를 빼고 CS50 자체 도구 외의 AI 소프트웨어로 답이나 코드를 제안·완성받는 것을 학업 부정행위로 보며 CS50 Duck을 포함한 자체 AI 소프트웨어 사용은 허용한다(CS50, Retrieved 2026). 2026년 가을 하버드 칼리지 실라버스는 학생 본인의 기말 프로젝트만 이 원칙의 예외로 두며 섹션을 소그룹 예약제 오피스아워로 보완한다. 퀴즈는 강의 시간에 강의실에서 대면으로 치르는 1시간 퀴즈 3회이며 각각 최종 성적의 15%를 차지한다(CS50, 2026).

CS50 Duck이 학생 메시지를 중계하는 Azure(Ⅰ장 3절)에 대해 Microsoft 문서는 Azure에서 판매하는 모델에 보낸 프롬프트·응답이 OpenAI 등 모델 제공자에게 제공되지 않는다고 설명한다. 프롬프트·응답은 고객의 허락이나 지시 없이 기반 모델 학습에 쓰이지 않는다. 같은 문서에 따르면 남용 모니터링에서 사람 검토용으로 프롬프트·응답을 저장하는 데이터 저장소가 있으며 현행 문서에는 보관 일수가 적혀 있지 않다(Microsoft, 2026). 이는 Azure 고객 일반을 대상으로 한 처리자 약정이어서 CS50.ai가 자체 서버에 남기는 대화 기록에는 적용되지 않고 CS50이 남용 모니터링에서 제외됐는지도 공개 자료로 확인되지 않는다.

CS50 팀은 학생과 CS50 Duck의 대화 메시지 1,000만 건을 분석해 2025년 논문에 활용했다(R. Liu et al., 2025). 대화 기록의 보관 기간·이용 목적·삭제 조항과 처리 국가는 CS50 공식 문서·실라버스에서 확인하지 못했다. cs50.ai의 이용 연령 기준, LTI 등 교육 데이터·콘텐츠 상호운용 표준의 지원, 교육 기술 상호운용 표준 단체 1EdTech의 인증도 문서로 확인되지 않았다. 프로필은 [표 Ⅲ-1]과 같다.

[표 Ⅲ-1] CS50 Duck 프로필

항목	내용
제품 유형·대상	하버드 CS50 강좌 팀이 만든 학생·교사용 ChatGPT 변형(CS50.ai), cs50.ai 웹과 CS50용 VS Code(cs50.dev)에 포함
교수법 설계	답을 곧바로 주기보다 학생을 답으로 이끄는 튜터 표방, 강좌별 규칙 시스템 프롬프트, 강의 자막 기반 RAG, 프롬프트 주입을 막는 요청 점검 '가드'
교수자·관리 통제	강좌 정책으로 CS50 자체 AI만 허용(2026년 가을, 기말 프로젝트 예외), 조교의 정기 모델 응답 비교 평가, 대면 퀴즈 3회(각 15%), 외부 기관 교수자용 설정·교사 자료는 문서로 확인되지 않음(CS50 강좌 내부 통제)
기반 모델·AI 공급	Azure 호스팅 GPT-4o(2025년 논문), 미세조정 결과 보고(2025년 논문), 현행 모델명 미확인, GitHub·Microsoft·OpenAI 지원
이용 형태·가격	CS50 강좌 학생·교사용(강좌 정책으로 과제에 허용), 별도 요금 안내는 공식 문서로 확인되지 않음
제공사 현황	Harvard CS50(대학 강좌 팀), 2026년 가을 강좌 실라버스에 허용 AI로 명시

항목	내용
데이터 처리	개발팀의 학생 대화 1,000 만 건 분석·논문 이용(이용 고지 미확인), 보관 기간·처리 국가·삭제 조항 미확인, 참고로 Azure 처리자 약정(모델 제공자 미제공·기반 모델 학습 미사용·남용 모니터링 저장)은 CS50 자체 기록에 적용되지 않음
연동·인증	LTI 등 표준 지원·1EdTech 인증 미확인
접근성·연령	이용 연령 기준 문서 미확인, 접근성 미측정(실측 결과 미수록)
독립 효과 근거	독립 혼합연구 1 편(학생 인식, 효과크기 없음), 참고로 개발팀 수행·자체 보고(학생 47% '매우 도움됨', 2023 년 여름 직원 검토에서 게시판 교과 답변 88% 정확, 가을 직원 승인 비율 39%는 저자가 과소평가로 해석)
이번 호 평가	자료 기준일까지 공개된 문서·연구로 판정(26 개 항목 중 12 개), 실측 필요 항목 14 개 미측정

자료: CS50·Microsoft 공식 문서, 개발팀 논문, 독립 연구를 바탕으로 연구진 정리

Duolingo 의 한국어 도움말은 Duolingo Max 가 Super 혜택에 더해 Explain My Answer(정답 해설)·Roleplay(롤플레이)·릴리와의 Video Call(영상 통화) 3 가지 AI 기능을 제공한다고 안내한다. 일부 AI 기능은 일부 과정에만 열려 있으며 Explain My Answer·Roleplay 제공 과정 목록에 한국어 사용자용 영어 과정은 없다(Ⅱ장 4절 바). Video Call 은 영어·프랑스어·스페인어·이탈리아어·독일어·포르투갈어를 배우는 Max 구독자만 쓸 수 있다고 안내한다(Duolingo, Retrieved 2026-a). 다른 공식 자료와는 두 군데가 어긋난다. 도움말은 Explain My Answer 를 Max 기능으로 소개하지만 2026 년 1 월 게시글은 원래 Duolingo Max 기능으로 시작한 Explain My Answer 를 모든 학습자에게 무료로 제공한다고 밝혔다(Mas Castillo, 2026). Video Call 의 대상 언어도 현행 블로그 본문은 일본어·한국어·중국어를 배우는 영어 사용자도 쓸 수 있다고 안내해 도움말과 다르다(Duolingo, 2023).

'듀오링고 교실'에 가입한 상태에서는 Duolingo Max 가 포함하는 Super 를 쓸 수 없고 교실을 탈퇴해야 구독을 살 수 있다. 구독은 취소하지 않으면 자동 갱신된다(Duolingo, Retrieved 2026-b). Duolingo 는 Duolingo for Schools 와 관련해 학교·교육구와 학생 데이터 프라이버시 협약(SDPA)을 체결하지 않는다고 밝힌다. 교사가 학생 계정을 직접 만드는 기능도 제거됐다(Duolingo, 2026g).

Duolingo 는 AI 기능 대화의 녹음·전사본을 저장해 자체 AI 모델의 학습·운영을 포함한 제품 개선·개인화에 쓸 수 있다고 밝힌다. AI 공급업체는 계약상 개인정보를 자체 목적으로 쓸 수 없고 개인정보는 일반적으로 계정 삭제 때까지 보관된다. 데이터는 미국에서 처리되며 미국 기준 13 세 미만이거나 각국 디지털 동의 연령 미만인 사용자는 아동 사용자로 따로 취급된다(Duolingo, 2026c). 약관은 학습 활동으로 생성된 데이터·자료(Activity Materials)를 Duolingo 소유로 정한다. 결제는 관할권 법률이 요구하는 경우를 빼고 환불되지 않으며 Duolingo 는 정확성이나 사용 결과로 얻을 성과를 보증하지 않는다(Duolingo, 2025).

Duolingo 는 2026 년 2 월 주주서한에서 Video Call 을 Duolingo Max 의 핵심 기능이라고 소개하면서 이를 Super 등급으로 옮기고 Max 의 구성 변경을 실험하겠다고 밝혔다(Duolingo, 2026a). 2026 년 8 월 주주서한에 따르면 신규 Super 구독자 대부분이 Video Call 을 쓸 수 있고 기존 Super 구독자에게는 연내 이용을 넓힐 예정이어서 Video Call 을 Max 구독자만 쓸 수 있다는 도움말 안내와 다르다(Duolingo, 2026d). 1EdTech 인증 등록부에서 Duolingo 제품의 인증 기록은 확인되지 않았다(1EdTech Consortium, Retrieved 2026). 프로필은 [표 Ⅲ-2]와 같다.

[표 Ⅲ-2] Duolingo Max 프로필

항목	내용
제품 유형·대상	언어 학습 앱 Duolingo 의 상위 구독 등급(Super 위), 개인 구독 학습자
교수법 설계	전문가가 쓴 Roleplay 시나리오·첫 메시지(과정 위치에 맞춤), 이전 통화를 기억하는 Video Call 캐릭터 릴리(2026. 8. 기준 신규 Super 구독자에게도 제공), Explain My Answer 의 오류 맞춤 해설(2026. 1. 무료화), 생성 내용 검토·부정확 응답 신고
교수자·관리 통제	교실 가입 중 Super 사용 불가(탈퇴해야 구독), Duolingo for Schools 신규 교실 불가·2027. 7. 31. 종료 예정, Max AI 기능의 교수자 개입 경로는 문서로 확인되지 않음
기반 모델·AI 공급	AI 기능 입력(텍스트·오디오)을 OpenAI·Google 같은 AI 공급업체와 공유 가능(계약상 자체 목적 사용 금지), 현행 모델명 미확인
이용 형태·가격	iOS·Android 앱 전용(웹 구매 불가), 일부 이용자에게 Super 보다 높은 가격(금액 미확인), 일부 과정만 제공, 자동 갱신, 관할권 예외를 빼고 비환불
제공사 현황	Duolingo(Form 10-K 연차보고서 공시), Video Call 의 Super 이동(2026. 2. 발표, 2026. 8. 신규 Super 구독자 대부분 제공)·Max 구성 변경 실험, Video Call 등 AI 비용 증가 공시(2025 회계연도)
데이터 처리	녹음·전사본 저장과 자체 모델 학습 활용 가능, 계정 삭제 때까지 보관, 학습 활동 자료 Duolingo 소유, 미국에서 처리
연동·인증	1EdTech 등록부 인증 기록 미확인, Schools 학생 데이터 프라이버시 협약(SDPA) 미체결
접근성·연령	13 세 미만·디지털 동의 연령 미만 아동 사용자 별도 취급, 접근성 미측정(실측 결과 미수록)
독립 효과 근거	Max 기능 단위 독립 효과 연구 없음(묶음 처치 독립 연구 2 편은 연구 질 문제로 제외), 일반 앱 독립 준실험 1 편(자기 보고 척도), 참고로 제공사 자체 보고(Video Call RCT 작은 정적 효과·동료심사 전, 자기효능감 조사·통제집단 없음)
이번 호 평가	자료 기준일까지 공개된 문서·공시·연구로 판정(26 개 항목 중 11 개), 실측 필요 항목 14 개와 A8 미측정

자료: Duolingo 공식 문서·공시, 1EdTech 등록부, 독립 연구를 바탕으로 연구진 정리

주: 데이터 처리 행의 근거 문서는 성격이 다르다(CS50 Duck 은 개발팀 논문과 Azure 처리자 약정, Duolingo Max 는 소비자용 개인정보처리방침과 약관).

2. 평가척도 적용 결과

두 제품은 자료 기준일까지 공개된 문서와 연구로 평가척도 26 개 항목을 검토했다. 실측 관찰이 있어야 판정하는 AUTO·HYBRID 항목(A1~A7·B2·B3·B5·B7·C1·C3·C6)은 이번 호에 실측 결과를 싣지 않아 판정하지 않고 미측정으로 두었으며 점수는 문서로 판정한 항목에서 나왔다. 두 제품은 양쪽 모두에서 판정한 공통 11 개 항목으로만 견줄 수 있다.

가. CS50 Duck 판정 근거

게이트 항목 B3(피드백·적응)·C1(타당성·신뢰성과 환각 관리)은 튜터 응답을 실측해야 판정할 수 있어 미측정이다. 두 항목을 판정하려면 수강생과 같은 조건에서 정답 직접 요구·오개념·허위 전제 같은 과업으로 튜터 응답을 실측해야 한다(Ⅳ 장 2 절 나). 참고 근거 가운데 정답을 곧바로 주지 않는 교수법 지시와 요청 점검 '가드'는 Ⅱ장 1 절 나에서 다뤘다. 개발팀 논문(자체 보고)에서 GPT-4o 를 쓴 기간만 보면 코드 블록 비율은 메시지 기준 25%로 전체 기간의 22%(Ⅱ 장 1 절 나)보다 높았고 대화 기준으로는 56%였다. 같은 논문의 평가 체계에는 조교가 정기적으로 검토하는 모델 평가 백엔드가

있으며 관리자가 평가 과제를 조교에게 배정하면 조교는 같은 학생 메시지에 대한 여러 모델의 응답을 비교해 선호하는 쪽을 고른다(R. Liu et al., 2025).

게이트 항목 C5(인간 통제·교수자 역할)는 부분 충족(◐)이다. 조교의 정기 비교 평가와 실라버스의 자체 AI 한정·대면 퀴즈(Ⅲ장 1절)가 확인되지만 AI 응답에 대한 사람의 관여는 모델 비교 평가로만 확인되고 지시 준수의 현재 수치도 공개 자료에서 확인되지 않았다. 기반 모델 갱신에 따른 튜터 행동 변화 위험(Ⅲ장 3절)도 남는다.

문서로 판정한 나머지 항목 가운데 학습목표 정렬(B1)과 동기·참여(B4)는 강좌 규칙·강의 자막에 맞춘 설계와 학생의 긍정적 평가가 확인되나 명시적 학습 목표가 문서로 확인되지 않고 과의존 위험도 함께 보고되어 부분 충족이다. 교수 통합·교수자 지원(B6)은 강좌 안에서 허용 AI 지정과 조교 평가로 통합되어 있으나 다른 기관 교수자가 CS50 Duck을 자기 강좌에 맞게 설정하는 방법이나 교사용 자료가 확인되지 않아 부분 충족이다. 표준 준수·상호운용(B8)은 LTI 등 표준 지원이나 인증이 확인되지 않아 미충족이다. AI 리터러시·연령 적합성(C8)은 공식 문서와 실라버스의 AI 사용 범위 안내가 학업 정직성 규정이어서 AI 리터러시 지원으로 보기 어렵고 이용 연령 기준과 학습자용 AI 리터러시 자료도 확인되지 않아 미충족이다.

프라이버시·데이터 거버넌스(C4)는 공식 문서·실라버스에 대화 기록의 보관 기간·이용 목적·삭제·연령 조항이 없고 1,000만 건 분석 이용에 관한 고지도 확인되지 않아 미충족이다. Azure 처리자 약정(Ⅲ장 1절)은 CS50의 고지를 대신하지 못한다. 책무성·거버넌스·정책 정합(C7)은 운영 주체가 분명하고 조교 정기 평가 체계가 있으나 이의 제기 절차·감사 기록과 국내 정책 대응이 확인되지 않아 부분 충족이다.

효과 근거로는 제공사와 독립된 뮌헨응용과학대 혼합연구가 있다. 설문·과제 자료에 따르면 학생들은 CS50 Duck의 비판단적 피드백을 높이 평가했고 정규 평가와 프로그래밍 교육에서 생성형 AI가 맡을 역할에 대해 단순하지 않은 견해를 보였다(Zönnchen et al., 2025). 이 연구는 초록 기준으로 통제집단과 학습 성과 효과크기를 보고하지 않아 효과 근거(B9·B10)는 미충족(○)이다. CS50 팀의 자체 조사에서는 학생 47%가 '매우 도움됨', 26%가 '도움됨', 21%가 '어느 정도 도움됨', 6%가 '도움 안 됨'으로 응답했다(R. Liu et al., 2024). 이 조사는 자기 보고 결과여서 효과 근거로 세지 않았다.

나. Duolingo Max 판정 근거

게이트 항목 B3·C1은 Max 기능의 응답을 실측해야 판정할 수 있어 미측정이다. 두 항목을 판정하려면 Max 기능이 제공되는 iOS·Android 앱 환경(Ⅰ장 3절)에서 튜터 응답을 실측해야 한다(Ⅳ장 2절 나). 참고 근거에는 제공사의 검토 체계 설명과 위험 공시가 함께 있다. Duolingo는 AI 생성 내용이 사실에 맞고 어조가 적절한지 계속 검토한다고 밝히며 학습자는 부정확한 AI 응답을 길게 눌러 신고할 수 있다고 안내한다(Duolingo, 2023). Duolingo는 연차보고서에서 생성형 AI가 부정확하거나 오도하는 내용과 환각을 낼 위험을 공시했다. 정확성 조치가 늘 성공하지는 않으며 오류 발견을 최종 이용자 등의 신고에 의존할 수 있다는 점도 밝혔다(Duolingo, 2026b).

게이트 항목 C5는 전문가가 쓴 Roleplay 시나리오·생성 내용 검토·이용자 신고 경로가 있으나 Duolingo Max AI 기능에 교수자가 개입하는 경로가 확인되지 않아 부분 충족(◐)이다. 교사는 Duolingo for Schools에서 학생의 친구 요청 주고받기 기능이나 성인 콘텐츠 표시를 켜거나 끌 수 있다(Duolingo, 2026g). 이 채널은 종료를 앞두고 있다(Ⅰ장 3절).

문서로 판정한 나머지 항목 가운데 교수 통합·교수자 지원(B6)은 교실 가입 중 Super를 쓸 수 없고 Duolingo for Schools가 종료될 예정이며 Max AI 기능을 수업에 통합하는 교사 자료가 확인되지 않아

미충족이다. 표준 준수·상호운용(B8)은 1EdTech 등록부의 인증 기록과 LTI 등 표준 연동 방법이 확인되지 않아(Ⅲ장 1절) 미충족이다. 프라이버시·데이터 거버넌스(C4)는 공급업체의 자체 목적 사용 금지와 접근·삭제 요청 경로가 있으나 녹음·전사본의 자체 모델 학습 이용과 계정 삭제 때까지의 보관이 확인되어 부분 충족이다. 학습목표 정렬(B1), 동기·참여(B4), 책무성·거버넌스(C7), AI 리터러시·연령 적합성(C8)도 부분 충족이고 주관적 만족도(A8)는 Max 기능 대상 공개 근거를 찾지 못해 미측정이다.

Duolingo Max 기능 단위로 효과를 가른 제공사 독립·고등교육 연구가 없어 효과 근거(B9·B10)는 미충족(○)이다. Max를 다른 앱·AI 도구와 묶어 처치한 독립 연구 2편은 Max의 효과를 따로 가를 수 없고 연구 질의 한계가 있어 반영하지 않았다. 제공사와 독립된 연구로는 중국 후난국제경제대학 외국어학과 1학년 80명을 대상으로 Duolingo가 의사소통 의지와 학습 참여에 미치는 영향을 본 준실험이 있다. 학습 참여 총점에서 Duolingo 사용의 효과크기는 Cohen's d = 2.03으로 보고되었지만 Max 기능이 아닌 일반 앱 사용을 다룬 결과이고 측정은 자기 보고 척도였다(Ouyang et al., 2024).

제공사 자체 보고로는 Duolingo 연구 보고서의 무작위 통제 실험(동료심사 전)이 있으며 B1.1 영어 과정의 일본어 화자 567명을 Video Call 조건과 통제 조건에 배정했다. 제공사 자체 보고에 따르면 Video Call 조건의 사후 말하기 점수는 회귀 분석에서 통제 조건보다 2점 높았고 보고서 스스로 이를 작은 정적 효과로 규정했다. 같은 보고서에서 Video Call 조건의 탈락률은 20.1%로 통제 조건 7.6%보다 높았다(Kittredge, Lee, & Jiang, 2025). 저자 6명 전원이 Duolingo에 재직한 자기효능감 연구에는 AI 기능을 쓰지 않은 통제집단이 없어 자기효능감 상승을 기능 때문이라고 단정할 수 없다(Kittredge, Hopman, et al., 2025). 두 자료는 효과 근거로 세지 않았다.

다. 점수와 판정

판정 규칙(Ⅰ장 4절)에 따라 문서 판정 항목은 두 제품 모두 자료 기준일까지 공개된 공식 문서와 연구로만 판정했고 효과 근거(B9·B10)에는 제공사와 독립적인 고등교육 연구만 반영했다. 실측이 필요한 항목(A1~A7·B2·B3·B5·B7·C1·C3·C6)은 실측 관찰이 있을 때만 판정한다는 규칙에 따라 모두 미측정으로 두었다. 게이트 B3·C1의 문서 근거는 부록 2 근거 칸에 참고로만 적었다. CS50 Duck은 튜터 설계 지향과 코드 블록 비율, 강의 자막 기반 응답이, Duolingo Max는 Explain My Answer 설명과 생성 내용 검토·신고 경로, 환각 위험 공시가 이에 해당한다.

투명성·설명가능성(C2)은 공식 문서의 기재로 판정했다. AI 기본법 시행령 제23조제1항 각 호도 사전 고지를 제품에 직접 기재하거나 계약서·사용 설명서·이용약관 등에 기재하는 방법과 이용자의 화면이나 단말기에 표시하는 방법 등으로 할 수 있게 정하고 있다(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령, 2026). 이 보고서는 시행령을 점검의 참고 기준으로만 썼으며 두 제품에 국내법이 적용되는지는 판단하지 않았다. CS50 공식 문서는 CS50.ai를 cs50.ai 웹과 CS50용 Visual Studio Code(cs50.dev)에 함께 들어 있는 학생·교사용 ChatGPT 변형으로 소개한다(CS50, Retrieved 2026). Duolingo 블로그의 현행 본문은 Duolingo Max가 Video Call과 Roleplay를 제공하며 두 기능이 생성형 AI 기술로 동작한다고 안내한다(Duolingo, 2023). 두 제품 모두 생성형 AI 기반임은 문서로 밝히지만 현행 기반 모델명과 응답 근거 제시 방식이 확인되지 않아 C2는 부분 충족(◐)이다.

축 점수는 소수 첫째 자리까지 반올림하고 종합 점수는 반올림한 축 점수로 계산했다. 측정 항목이 없는 축은 점수를 내지 않았고 종합은 점수가 있는 축의 가중치 합으로 나누어 다시 맞췄다(재정규화). 결과는 [표 Ⅲ-3]과 같다.

[표 Ⅲ-3] 평가 결과 요약

구분	CS50 Duck(Harvard CS50)	Duolingo Max(Duolingo)
판정 근거	자료 기준일까지 공개된 공식 문서·연구(문서 판정)	자료 기준일까지 공개된 공식 문서·공시·연구(문서 판정)
측정 범위	12/26 항목(A 1·B 6·C 5), 필수 5/13	11/26 항목(A 0·B 6·C 5), 필수 5/13
미측정 항목	14 개(A1~A7·B2·B3·B5·B7·C1·C3·C6, 모두 실측 필요 항목)	15 개(CS50 Duck 과 같은 14 개와 A8)
축 A 사용성(가중치 30)	50.0(A8 1 개 항목, 문서 판정)	미산출(측정 항목 없음)
축 B 교육적 특성(가중치 40)	26.9	19.2
축 C AI 신뢰성·윤리·거버넌스(가중치 30)	33.3	50.0
종합 점수	35.8(축 A·B·C)	32.4(축 B·C 로 재정규화)
공통 11 개 항목 기준	29.6	32.4
게이트 항목(B3·C1·C5)	B3 미측정·C1 미측정·C5 ◐	B3 미측정·C1 미측정·C5 ◐
고부담 게이트	해당 없음	해당 없음
도입 검토 의견	판정 유보	판정 유보
핵심 보완점	수강생과 같은 조건의 튜터 응답 실측(B3·C1), 현행 모델의 지시 준수 수치와 사람의 검토 경로, 대화 기록 보관·이용 조항, 학습자용 AI 리터러시·이용 연령 안내	모바일 앱 환경의 Max 기능 실측(B3·C1), 교수자 개입 경로, Max 기능 단위 독립 효과 연구

자료: 공식 문서·연구 판정을 바탕으로 연구진 정리

주: ◐는 부분 충족이며 미측정 항목은 점수 계산에서 제외했다. 고부담 게이트는 이번 호 솔루션군에 적용하지 않았다. 항목별 판정과 근거는 부록 2 에 실었다.

두 제품 모두 필수 항목 13 개 가운데 5 개(B1 학습목표 정렬·B9·C2·C4·C5)만 판정되어 경계값을 적용할 최소 범위(10 개)에 못 미쳤으므로 종합 점수와 관계없이 '판정 유보'다. 판정 유보는 결함 판정이 아니라 측정 범위가 부족하다는 뜻이다. 미측정 항목을 모두 ●로 두면 종합은 CS50 Duck 69.1 점, Duolingo Max 71.9 점(조건부 검토)이고 모두 ○로 두면 각각 14.0 점, 13.6 점(신중 검토)이다. 게이트 항목 C5 가 두 제품 모두 ◐이고 필수 항목 B9(두 제품)와 C4(CS50 Duck)가 ○여서 이 판정들이 바뀌지 않는 한 대상 기능을 실측해도 의견의 상한은 조건부 검토다.

CS50 Duck 의 종합 35.8 점은 축 A·B·C(가중치 합 100)로, Duolingo Max 의 종합 32.4 점은 축 B·C(가중치 합 70)로 계산해 항목 구성이 다르므로 두 종합 점수는 직접 비교하지 않는다. CS50 Duck 의 축 A 50.0 점은 주관적 만족도(A8) 한 항목을 독립 연구의 학생 평가로 판정한 값이며 사용성 실측 결과가 아니다. Duolingo Max 는 같은 기준을 충족하는 근거를 찾지 못해 A8 이 미측정이다. 두 제품을 견줄 때는 공통 11 개 항목(B1·B4·B6·B8·B9·B10·C2·C4·C5·C7·C8)의 기준 점수(CS50 Duck 29.6 점, Duolingo Max 32.4 점)만 쓴다. 판정이 다른 항목은 B6(교수 통합·교수자 지원: CS50 Duck ◐, Duolingo Max ○), C4(프라이버시·데이터 거버넌스: CS50 Duck ○, Duolingo Max ◐), C8(AI 리터러시·연령 적합성: CS50 Duck ○, Duolingo Max ◐) 3 개다. 2.8 점 차는 이 차이들이 상쇄되고 남은 값이다.

공통 항목은 모두 문서 판정 항목이어서 정답 보류·오개념 교정·환각 대응·사용성의 실측 결과가 들어 있지 않다. 두 제품의 B9 미충족도 제공사 독립·고등교육 대상·제품 단위 효과 연구가 없다는 근거 부재이지 효과가 없다는 뜻은 아니다. CS50 Duck 의 C8 미충족도 자료 기준일까지 공개된 문서에서 이용 연령 기준과 학습자용 AI 리터러시 자료를 확인하지 못했다는 뜻이다. 이 점수들은 판정 범위가 다른 문서 기반 진단 정보이며 순위로 읽을 수 없다. 항목별 판정과 근거는 부록 2 에 실었다.

3. 비교 종합

두 제품은 같은 'AI 튜터'라는 이름 아래 공급 형태와 통제 구조가 다르다. CS50 강좌는 다른 사람과 AI 와의 상호작용이 강좌 내용 숙달을 도울 수 있다고 인정한다(CS50, 2026). 그러면서 강좌 정책으로 허용 AI 를 자체 도구로 한정하고 조교가 모델 응답을 정기적으로 비교 평가한다(Ⅲ장 1 절, Ⅲ장 2 절 가). 공식 문서는 CS50.ai 를 학생·교사용 도구로 소개하지만 다른 기관 교수자가 CS50 Duck 을 자기 강좌에 맞게 설정하는 방법은 공개 문서로 확인되지 않았다. 문서로 보면 Duolingo Max 의 교수법은 제공사가 통제한다. 전문가가 쓴 Roleplay 시나리오와 생성 내용 검토·이용자 신고 경로가 있으나 교수자가 AI 기능에 개입하는 경로는 문서로 확인되지 않았다(Ⅱ장 1 절 마, Ⅲ장 2 절 나). Duolingo Max 는 개인이 구독하는 등급이며 교실을 탈퇴해야 Super 구독을 살 수 있다. 여러 구독의 일괄 구매는 고용주가 직원용으로 할인 구매하는 Duolingo for Business 로 안내된다(Duolingo, Retrieved 2026-b).

운영 위험의 성격도 다르다. CS50 개발팀은 OpenAI 의 기반 모델 갱신이 튜터 행동을 예기치 않게 바꿀 수 있다고 보고 GPT-4 에서 GPT-4o 로 바꾼 전환을 그 예로 들었다(R. Liu et al., 2025). Duolingo Max 는 Explain My Answer 무료화(2026 년 1 월)와 Video Call 의 Super 등급 이동처럼 기능이 구독 등급 사이를 옮겨 가고 최상위 등급의 구성 변경 실험도 예고되었다(Ⅲ장 1 절).

이 비교는 실측 결과 없이 자료 기준일까지 공개된 문서로 견준 것이어서 순위를 매기지 않는다. 상대 제품 문서에서 확인되지 않은 항목도 기능이 없다고 단정할 수는 없다. 대학은 대학 자체 강좌 튜터와 개인 구독형 언어 튜터를 [표 Ⅲ-4] 마지막 열의 같은 질문으로 확인할 수 있다.

[표 Ⅲ-4] 대표 솔루션 비교

비교 차원	CS50 Duck	Duolingo Max	대학 도입 시 확인할 점
공급 형태	대학 강좌 팀이 만든 강좌 전용 튜터(cs50.ai 웹·CS50 용 VS Code)	B2C 앱의 상위 구독 등급(모바일 앱 전용, 자동 갱신)	운영 주체와 비용 부담(자체 개발·상용 구독)
대학 도입 경로(문서)	학생·교사용 소개, 외부 기관의 자기 강좌용 CS50 Duck 설정 방법 미확인	개인 구독(앱 전용, 일부 이용자에게 Super 보다 높은 가격, 금액 미확인), 교육기관용 Max 안내 미확인·Schools 종료 예정, 일괄 구매는 고용주용 Duolingo for Business 로 안내	외부 기관의 도입·설정 경로와 비용 부담 주체
교수법 설계(문서)	답으로 이끄는 튜터 설계, 강의 자막 기반 응답, 요청 점검 '가드'	전문가가 쓴 Roleplay 시나리오, 이전 통화를 기억하는 Video Call 캐릭터, 오류 맞춤 해설	정답 보류·오개념 교정의 실측 결과
정답 보류·내용 정확성(실측)	미측정(B3·C1, 실측 결과 미수록)	미측정(B3·C1, 실측 결과 미수록)	수강생·기관 계정으로 시범 측정

비교 차원	CS50 Duck	Duolingo Max	대학 도입 시 확인할 점
문서상 위험 신호	GPT-4o 기간(2025 년 논문) 코드 블록 응답 메시지 25%·대화 56%(개발팀 수행, 현재 수치 미확인)	환각 위험·이용자 신고 의존 공시	현재 값의 공개와 재점검 주기
교수자 통제(문서)	강좌 정책(자체 AI 만 허용), 조교의 모델 응답 평가, 외부 기관 교수자용 설정·교사 자료 문서 미확인(CS50 강좌 내부 통제)	교실 가입 중 Super 사용 불가, Duolingo for Schools 종료 예정, Max AI 기능의 교수자 개입 경로 문서 미확인	교수자 관리 경로와 설정 범위
데이터 조건(문서)	개발팀의 대화 1,000 만 건 분석·논문 이용(이용 고지 미확인), 보관 기간·처리 국가 미확인, Azure 처리자 약정(모델 제공자 미제공·학습 미사용·남용 모니터링 저장)은 CS50 자체 기록에 미적용	녹음·전사본의 자체 모델 학습 이용 가능, 계정 삭제 때까지 보관, 미국에서 처리	보관 기간·2 차 이용·국외 처리
운영 위험	기반 모델 갱신에 따른 행동 변화	등급 구성 변화(Explain My Answer 무료화, Video Call 의 Super 이동, Max 구성 변경 실험)	서비스 연속성과 가격 변화 대응
국내 이용	국내 도입 실적·외부 기관 도입 경로 미확인, 이용 연령 기준 문서 미확인	한국어 도움말·약관 제공, Explain My Answer·Roleplay 제공 과정에 한국어 사용자용 영어 과정 없음	한국어 학습자 지원 범위
효과 근거(독립)	독립 혼합연구 1 편(학생 인식), 효과크기 없음	Max 기능 단위 독립 연구 없음(묶음 처치 독립 연구 2 편 제외, 일반 앱 준실험·제공사 자체 보고만)	학내 효과 측정 계획
접근성(A6·A7)	미측정	미측정	대상 기능 화면의 접근성 점검

자료: 본문 Ⅰ~Ⅲ장의 인용 자료를 바탕으로 연구진 정리

주: 문서상 위험 신호 행은 공개된 자료의 종류가 달라(CS50 Duck 은 개발팀 측정값, Duolingo Max 는 위험 요인 공시) 위험 크기의 비교가 아니다. 미측정은 이번 호에 실측 결과를 싣지 않았다는 뜻이다.

Ⅳ. 시사점 및 종합

1. 국내 적용 시사점

Ⅱ·Ⅲ 장의 분석에 비추어 AI 튜터를 국내 대학 강좌에 들일 때의 법제·데이터 보호·대학 정책·기관 도입 모델 시사점은 [표 Ⅳ-1]에 수준별로 요약했다.

[표 Ⅳ-1] 수준별 시사점

수준	확인된 사실(요지)	시사점	점검 질문
국가	AI 기본법 시행령의 고영향·생성형 AI 사전 고지(2026. 1. 22. 시행), 저작권법의 수업 목적 이용 요건(보상금·복제방지조치)과 프로그램 특례, 행동계획의 학생 개인정보 보호 지침 권고(2026 년 4 분기까지)	학내 튜터의 고지 방식과 강좌 자료 이용 범위를 대학이 스스로 정해야 할 수 있어 고지 적용 범위·강좌 자료 연결에 대한 정부 해석 안내가 대학별 판단 부담을 줄일 수 있음	화면 고지가 있는가, 강좌 자료의 이용 허락 범위를 확인했는가
기관	처리 위탁의 문서화·감독 의무, 국외 이전의 원칙적 금지, 대학 지침의 책임 개인 전가 경향	대화 기록의 위탁·국외 처리·보관 조건을 계약으로 관리	보관 기간·2 차 이용·처리 국가를 계약서로 확인했는가
교수자·조교	학생 교육·지도는 교원, 조교는 교육·연구·학사 사무 보조, 조교·강사 인력 축소 우려(언론)	AI 는 반복 질문, 조교는 응답 검수·심화 지도를 맡는 역할 재설계	조교의 응답 검수를 공식 업무로 정했는가
학습자	한국어 성능 격차, 선택형 튜터의 낮은 사용률, 인간 교수자에게 질문을 망설인 학생 58%와 이들의 더 잦은 VTA 사용(KAIST 강좌, 개발팀 수행)	한국어 실측, 사용 목적 안내, 질문을 망설이는 학생의 질문 창구(개발팀 수행, 집단 비교)	한국어 강좌 자료로 시범 운영했는가

자료: 본문 Ⅱ·Ⅲ·Ⅳ 장 인용 자료를 바탕으로 연구진 정리

가. 법제

AI 기본법 시행령 제 23 조제 1 항은 고영향·생성형 AI 를 이용한 제품·서비스를 제공하는 인공지능사업자에게 해당 AI 에 기반해 운용된다는 사실을 이용자에게 사전에 고지하도록 정했다(Ⅰ 장 1 절 가). 사전 고지는 제품등에 직접 기재하거나 계약서·사용 설명서·이용약관 등에 기재하는 방법, 이용자의 화면이나 단말기 등에 표시하는 방법, 제공 장소에 게시하는 방법, 과학기술정보통신부장관이 인정하는 방법 가운데 하나로 해야 한다. 제품·서비스명이나 화면 문구로 AI 활용 사실이 명백하거나 인공지능사업자의 내부 업무 용도로만 사용되는 경우는 적용 예외가 될 수 있다(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령, 2026). 주된 이용자의 나이와 신체적·사회적 조건 등을 고려해 명확하게 인식할 수 있게 하라는 기준(제 23 조제 3 항)은 법 제 31 조제 3 항에 따른 고지·표시에 적용된다(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령, 2026). 법 제 31 조제 3 항은 실제와 구분하기 어려운 가상의 음향·이미지·영상 등 결과물의 고지·표시를 규정한다(인공지능 발전과 신뢰 기반 조성 등에 관한 기본법, 2026).

저작권법 제 25 조제 3 항은 고등교육법에 따른 학교 등이 수업 목적으로 이용하는 경우 공표된 저작물의 일부분을 복제·공중송신할 수 있게 한다. 저작물을 이용하려는 자는 문화체육관광부장관이 고시하는 기준의 보상금을 지급해야 하고 학교가 공중송신할 때는 복제방지조치 등 대통령령으로

정한 조치를 해야 한다. 컴퓨터프로그램에는 이 조항이 적용되지 않으며 제 101 조의 3 에 따라 학교에서 교육을 담당하는 자는 수업과정에 제공할 목적으로 필요한 범위에서 공표된 프로그램을 복제·배포할 수 있다(저작권법, 2026). 시행령 제 9 조가 정한 필요한 조치에는 전송하는 저작물을 수업을 받는 자 외에는 이용할 수 없도록 하는 접근제한조치와 복제방지조치가 포함된다(저작권법 시행령, 2026).

문화체육관광부·한국저작권위원회 안내서는 RAG 시스템의 설계 방식과 외부 데이터베이스 저작물의 노출 정도에 따라 침해 판단 가능성이 달라질 수 있다고 보며 외부 데이터를 쓰려면 이용 허락이 필요할 수도 있다고 안내한다. 이 안내서에 따르면 2025 년 6 월 기준 권리자 허락 없는 AI 학습 활용이 저작권 침해인지에 대한 법원 판단은 없었다(문화체육관광부·한국저작권위원회, 2025).

이 조문들만으로는 대학이 학내 튜터를 학생에게 제공할 때 '인공지능사업자'에 해당하는지, 강좌 자료를 튜터에 연결해 외부 사업자 서버로 보내는 일이 수업 목적의 복제·공중송신 범위에 드는지가 분명하지 않다(연구진 해석). 예제·해설 코드를 다루는 코딩 튜터는 프로그램 규정을 따로 살펴야 할 수도 있다. 대학은 화면 고지 문구, 강좌 자료의 이용 허락 범위, 수강생 외 접근 제한을 도입 전에 점검해 둘 만하다(부록 4). 관계 부처가 대학 학내 튜터의 사전 고지 적용 범위와 강좌 자료의 RAG 연결에 관한 해석 안내를 마련하면 대학마다 따로 판단해야 하는 부담을 줄일 수 있다(연구진 해석).

나. 데이터 보호

개인정보 보호법상 개인정보처리자가 개인정보 처리 업무를 제 3 자에게 위탁하려면 위탁업무 수행 목적 외 처리 금지 등 법정 내용을 담은 문서로 해야 한다. 위탁자는 수탁자를 교육하고 수탁자가 개인정보를 안전하게 처리하는지 감독해야 한다. 개인정보처리자는 원칙적으로 개인정보를 국외로 제공·처리위탁·보관해서는 안 되며 법이 정한 경우에만 국외로 이전할 수 있다(개인정보 보호법, 2026).

KAIST 가상 조교를 개발해 평가한 연구팀에 따르면 대화 로그에는 학번·대화 이력·질문·응답·시각·검색 문서가 함께 기록되었다(Kweon et al., 2025). 국내 한 대학의 파이썬 입문 강좌 튜터(Ⅱ 장 4 절 바)는 학생·교수자 데이터에 관한 대학 개인정보 정책상 폐쇄형 상용 모델과 학생·교수자의 직접 실시간 상호작용이 허용되지 않아 배치 모델로 오픈소스 모델을 골랐다(J. Lee et al., 2026). 대표 솔루션 가운데 Duolingo 는 개인정보 처리방침에서 데이터를 미국에서 처리하며 미국이 이용자 거주 관할권과 같은 수준의 데이터 보호를 제공하지 않을 수 있다고 고지한다(Duolingo, 2026c).

12 개국 43 개 대학의 생성형 AI 지침을 프라이버시·보안 관점에서 분석한 연구에서 IT 부서 지침 19 곳 중 16 곳은 기존 개인정보 보호 정책을 다시 보라고만 안내했다(Ng et al., 2026). 국내에서는 국가인공지능전략위원회와 과학기술정보통신부가 2026 년 2 월 「대한민국 인공지능행동계획(인공지능 기본계획 2026~2028)」을 확정했다. 행동계획은 교육부가 개인정보보호위원회와 협의해 고등교육법에 따른 학교를 포함한 각급 학교의 교육 AI 개발·활용 시 학생 개인정보 보호 지침을 2026 년 4 분기까지 마련하도록 권고했다(국가인공지능전략위원회, 2026).

학번이 함께 남는 대화 로그는 개인정보로 다뤄야 할 것으로 보인다. 외부 API 나 해외 처리를 쓰는 튜터라면 위탁과 국외 이전 요건을 계약 단계에서 확인해야 한다. 대표 솔루션 두 제품도 대화 기록의 보관 기간과 2 차 이용 조건을 문서로 밝힌 정도가 달랐다(Ⅲ 장 3 절). CS50 Duck 은 처리 국가와 보관 기간을 문서로 밝히지 않았고 Duolingo 는 미국 처리와 계정 삭제 때까지의 보관을 고지했다(Ⅲ 장 1 절). 대학 지침이 기존 정책 안내에 머물면 판단 부담이 개인에게 넘어갈 수 있어 기관 차원의 계약 점검과 대화 로그의 보관·가명처리 기준을 마련해 둘 필요가 있다(부록 4).

다. 대학 정책

고등교육법상 학생을 교육·지도하는 임무는 교원에게 부여돼 있다. 조교의 임무는 교육·연구 및 학사에 관한 사무의 보조이고 조교가 될 수 있는 사람의 자격기준은 대통령령으로 정하도록 돼 있다(고등교육법, 2026).

언론 보도에서 KAIST 관계자는 AI가 반복적인 기본 질문을 먼저 처리해 조교가 복잡한 질문과 심화 설명에 집중할 수 있다고 말했고 일부 전문가는 AI 도입이 조교·강사 인력 축소로 이어질 수 있다고 우려했다. 기사는 대다수 교육 정책 전문가들이 AI 도입 자체보다 운영 방식이 중요하다고 본다고 전했다(김의진, 2026).

국내 원격대학 재학생 면담 연구(질적 연구)에서 학습자들은 AI 튜터에 학사 운영·진도 관리 같은 관리 역할과 즉각적 응답을 기대했지만 정보 신뢰성과 정서적 상호작용의 한계를 우려했다(정재현 외, 2025). OECD는 동기·관계·사회정서 학습은 본질적으로 인간의 책임으로 남는다고 보았다(OECD, 2026a). 한국어 의료윤리 문항 연구(Ⅱ장 4절 바)의 저자들은 언어별 AI 성능 차이가 학습자에게 불평등한 혜택을 낳을 수 있다고 지적했다(Kim et al., 2025).

'AI 조교'라는 이름의 서비스는 자격기준이 법령으로 정해지는 법률상 조교와 다른 것으로 읽히며 교원의 교육·지도 임무를 AI가 대신할 근거 규정은 조문에서 확인되지 않는다(조문에 기댄 연구진 해석). 대학은 AI가 맡을 반복 질문·즉각 응답과 사람이 맡을 응답 검수·심화 지도·관계 형성을 나누는 운영 기준으로 정책을 정리할 수 있다. 한국어로 운영하는 강좌는 도입 전에 한국어 자료와 질문으로 응답 품질을 실측해야 한다.

라. 기관 도입 모델

대학 자체 개발형은 API 운영비가 적게 든 사례가 있지만(KAIST 대학원 강좌 1개, 14주 약 180달러, Ⅱ장 2절 다) 개발과 검수에는 인력이 든다. 하버드 물리 RCT를 수행한 튜터 개발팀은 수업 준비에서 가장 많은 시간이 교수법 원칙을 반영한 AI 튜터 플랫폼 소프트웨어 개발에 들었다고 밝혔다(Kestin et al., 2025). 검수 인력 면에서는 국내 한 대학 파이썬 강좌(Ⅱ장 3절 마)의 조교가 학생을 오도하거나 지나치게 긴 사전 생성 질문·답변을 걸러내는 검수 역할을 맡았다(J. Lee et al., 2026).

교수법 통제 주체 면에서 상용 플랫폼형인 엘리스 AI 헬피는 교육자가 작성한 수업 자료 지도서와 지시사항을 기반으로 답변을 생성하고(Ⅱ장 1절 가) 채팅 내역은 자동으로 저장된다. 엘리스는 교육자 매뉴얼에서 지도서가 충실할수록 학생이 받는 안내가 더 정확해진다고 안내한다(엘리스, Retrieved 2026). AUT 사례(Ⅱ장 2절 라)의 교수자는 중요한 평가 기간의 시스템 접속 불가를 지적했고 시차 때문에 공급사 지원 전에 문제를 먼저 파악해야 하는 경우가 많았다. 피드백 비일관성과 언어 적응 한계, 기관 통합 문제도 AI 조교 도입의 긴장 요인으로 나타났다(Gedera & Griffiths, 2026).

교수자 설계 맞춤형의 예로 국내 대학 교양 일본어 강좌의 교수자는 주교재 PDF와 어휘·문형 파일을 GPTs에 올려 지식 기반을 구성했다. 연구팀은 학습자의 정서적 안전감이 기술 자체보다 교수자가 설계한 비판단적·지지적 상호작용 환경에서 비롯된 것으로 해석했다(H. Lee & Kwon, 2026). 인간 튜터 결합형으로는 대규모 강좌의 피드백·상호작용 제한을 보완하는 1:1 면담 중심 첨삭 체계인 대학 글쓰기 전문 튜터 제도에 생성형 AI 활용을 결합한 사례가 있으며 이 사례는 통제집단 없이 평가되었다(김혜림·정향윤, 2026).

개인 구독형의 이용 조건은 제공사 약관과 등급 구성을 따른다. 공정거래위원회는 결제일로부터 30일이 지나면 환불을 해 주지 않던 Speak의 환불 제한 약관 조항이 약관법에 위반된다고 판단했다(

공정거래위원회, 2024). Duolingo 도 Video Call 을 Super 등급으로 옮기고 최상위 구독 등급 Duolingo Max 의 구성 변경을 실험하겠다고 밝혀(제공사 발표, Ⅱ 장 2 절 나, Ⅲ 장 1 절) 개인 구독형에서는 제공사가 등급 구성을 바꿀 수 있다(Duolingo, 2026a).

다섯 유형은 교수법을 통제하는 주체, 응답을 검수할 인력, 데이터 계약, 서비스 지속성에서 차이가 난다([표 Ⅳ-2]). 자체 개발형은 운영비보다 개발·검수 인력이, 상용 플랫폼형은 LMS 연동과 평가 기간 지원이 성패를 좌우할 가능성이 있다. 개인 구독형을 강좌에 쓰려면 등급·가격이 바뀔 수 있고 비용이 학생 개인에게 넘어간다는 점(개인 구독·교육기관용 안내 미확인, Ⅲ 장 3 절)을 전제로 해야 한다.

범용 LLM 을 이미 기관 단위로 제공하는 대학은 전용 튜터를 따로 들일지 판단할 때 강좌 자료에 근거한 응답, 응답 점검·순차 안내 같은 정답 보류의 구조적 강제, 교수자 통제와 조교 검수 경로, 데이터 계약이 실제로 더 확보되는지를 기준으로 삼을 수 있다(연구진 해석, 부록 4). 범용 LLM 학습 모드는 내부 구조가 공개되지 않아 이번 호에서 직접 비교하지 못했고(Ⅰ 장 3 절) 학습 효과를 견준 대학 실험도 학습 모드가 아닌 범용 채팅과의 비교였다(Ⅱ 장 3 절 나).

[표 Ⅳ-2] AI 튜터 기관 도입 모델 비교

점검 차원	대학 자체 개발형 (CS50 Duck·KAIST VTA·SCALA)	상용 플랫폼형(엘리스 AI 헬피 ·AUT NF AI)	교수자 설계 맞춤형(교양 일본어 GPTs)	개인 구독형 (Duolingo Max·Speak)	인간 튜터 결합형(대학 글쓰기 전문 튜터)
교수법 통제 주체	대학(강좌 정책·시스템 지시·조교의 모델 평가)	교육자가 쓰는 지도서·지시사항	교수자(지식 기반 ·상호작용 설계)	제공사(전문가 시나리오·생성 내용 검토, Duolingo)	전문 튜터(AI 결과물 정제·재구성 지도)
비용·운영	API 운영비 사례 (KAIST 대학원 강좌 1 개, 14 주 약 180 달러), 수업 준비 시간 대부분이 플랫폼 개발(하버드 물리 튜터)	기관 관리자 활성화(엘리스), LMS 연동은 한 강좌만(AUT)	확인 자료 없음	개인 구독(Speak 월간 29,000 원 등, 2024. 12. 기본가), 교육기관용 안내 미확인(Duolingo)	1:1 면담 운영 (비용 자료 없음)
데이터	학번 포함 대화 로그(KAIST), 대학 정책에 따른 오픈소스 모델 (SCALA)	채팅 내역 자동 저장(엘리스)	교재 파일의 외부 플랫폼 업로드	녹음·전사본 자체 모델 학습 가능, 미국 처리 (Duolingo)	확인 자료 없음
교수자·조교 역할	성적 관련 질문은 사람(KAIST), 조교 검수(SCALA)·모델 응답 비교(CS50)	지도서 작성, 공급사 지원 전 문제 파악(AUT)	설계·운영	교수자 개입 경로 확인되지 않음 (Duolingo Max)	1:1 면담 첨삭
서비스 지속성	기반 모델 갱신에 따른 행동 변화 (CS50)	평가 기간 접속 불가(AUT)	확인 자료 없음	Duolingo for Schools 종료 예정(2027. 7. 31.)·등급 구성 변경(Duolingo)	확인 자료 없음

점검 차원	대학 자체 개발형 (CS50 Duck·KAIST VTA·SCALA)	상용 플랫폼형(엘리스 AI 헬피 ·AUT NF AI)	교수자 설계 맞춤형(교양 일본어 GPTs)	개인 구독형 (Duolingo Max·Speak)	인간 튜터 결합형(대학 글쓰기 전문 튜터)
확인된 위험	지시 희석(CS50), 교수진보다 낮은 신뢰(KAIST)	피드백 비일관성, 언어 적응 한계 (AUT)	범용 GPT 와의 비교 검증 없음	환불 제한 약관 (Speak, 시정), 한국어 사용자용 영어 과정이 Explain My Answer·Roleplay 제공 과정 목록에 없음(Duolingo 도움말)	통제집단 없는 평가

자료: 본문 Ⅱ·Ⅲ 장과 Ⅳ장 1 절 인용 자료를 바탕으로 연구진 정리

주: '확인 자료 없음'은 인용 자료에 해당 정보가 없다는 뜻이며 문제가 없다는 뜻이 아니다. 본교 서비스는 이해관계 때문에 이 비교에 넣지 않았다(Ⅳ 장 2 절 가).

2. 본교·연구회 활용 제언

가. 본교 적용 사례(관계 공개)

이 보고서는 HAI-GPT 를 운영하는 본교 AI 에듀테크센터의 연구과제 산출물이며, 연구진에는 해당 부서·센터 구성원이 포함된다. 이 센터가 운영·실증하는 HAI-MathTutor(AI 선형대수 튜터)·HAI-Tutor·K-University 'AI 동반자'와 HAI-GPT 스터디 튜터는 이해관계 때문에 채점하지 않고 공개된 사실만 적는다. 대학 발표의 'AI 선형대수 튜터'는 센터 서비스 HAI-MathTutor 와 접속 주소가 같아 이 보고서에서는 하나의 서비스로 다룬다. 이 보고서가 게시되는 KELI 플랫폼의 솔루션 마켓에는 센터가 운영하는 HAI-MathTutor(솔루션 공급 기업 슈퍼러닝)가 함께 소개되어 있다.

본교의 '한림 AI 교육 솔루션'은 교수·학습·진로지도 기능을 통합 지원하는 플랫폼이며 그 지원 수단으로 AI 튜터·AI 조교·AI 도우미·AI 교수로 구성된 AI 동반자(Companion)를 둔다(본교 제공 자료를 바탕으로 한 언론 기획 기사). 본교는 2025 년 초 AI 조교·AI 도우미 1 차 서비스를 시작했고 AI 조교는 강의자료 작성과 학생별 맞춤 채점·질문 답변·상담 지원 등 교수 업무를 지원하는 서비스다(김소현, 2025). KAIST 가상 조교의 응답 생성 프롬프트는 VTA 의 임무를 강좌 관련 질문에 답하며 학생을 돕는 것으로 정의한다(Kweon et al., 2025). 같은 '조교'라는 이름이라도 두 서비스의 지원 대상은 교수자와 학생으로 다르다.

본교는 에듀테크 기업 슈퍼러닝과 함께 개발한 'AI 선형대수 튜터'를 2025 년 가을 학기부터 정식 도입했다(본교 발표). 이 튜터는 손글씨 풀이를 분석해 계산 실수와 개념 오해를 진단하며 학생이 어려움을 겪으면 정답 대신 단계별 힌트를 주도록 설계됐다(한림대학교, 2025). KELI 솔루션 마켓은 HAI-MathTutor 를 개인별 맞춤형 학습을 지원하는 AI 수학 튜터로 소개하며 솔루션 공급 기업을 슈퍼러닝으로 적는다(한림대학교 KELI 사무국, Retrieved 2026). 2026 학년도 2 학기 AI-HiED 교과목 운영을 위해 교수자와 AI 교육서비스를 연결한 매칭데이에서는 HAI-Tutor·HAI-MathTutor 등 HAI 에코시스템을 구성하는 서비스 20 개가 소개됐고 교수자 교육 수요·서비스별 지원 범위·기업 운영 여건이 검토됐다(한림대학교, 2026).

공개 자료에서는 이들 서비스의 학습 효과나 2 학기 교과목 적용 결과를 검증한 연구·평가를 찾지 못했다. AI 선형대수 튜터의 오류 진단·단계별 힌트 설명은 대학 자기 발표이며 이번 호는 이 기능이

설계대로 작동하는지 확인하지 않았다. 채점하지 않은 서비스에 결함 근거가 드러나지 않은 것을 상대적 우위로 읽어서는 안 된다. IV 장 2 절 나의 역할 분담·실측·데이터 점검 제언은 학내 서비스에도 같은 기준으로 적용한다. II 장 4 절 마의 의인화·정서 기능 점검 기준도 'AI 동반자'에 똑같이 적용한다 .

나. 제언

교무 부서와 교수학습개발센터는 반복·기본 질문과 즉시 응답은 AI 가, 성적 관련 질문과 응답 검수·심화 지도는 교원·조교가 맡도록 역할을 미리 나누고 조교의 검수를 공식 업무로 정하는 방안을 검토할 수 있다(IV 장 1 절 다). KAIST 가상 조교 강좌도 성적 관련·중요 질문은 실시간 수업이나 Blackboard 질의응답으로 처리하고 일반 내용·코딩 질문에는 VTA 를 쓰도록 권했다(Kweon et al., 2025).

교수학습개발센터와 AI 에듀테크센터는 도입 전에 정답 직접 요구·개념 설명·이해 점검·오개념·허위 전제·근거 요청 같은 과업을 한국어 자료로 실행해 정답 보류와 내용 정확성을 실측하고 기반 모델이 바뀔 때마다 다시 점검할 수 있다. 공개 벤치마크는 모델을 고르는 참고로 쓸 만하다. Maurya 외(2025)는 7 개 LLM·인간 튜터의 192 개 대화와 1,596 개 응답에 8 개 교수 차원별 기준(골드) 주석을 단 MRBench 를 공개했다. Macina 외(2025)는 자신들의 벤치마크가 고등학교 수학 다단계 문제에 한정되며 학습 성과를 측정하는 인간 연구를 대체하려는 것이 아니라 어떤 모델을 쓰고 비교할지 판단하는 척도를 목표로 한다고 밝혔다.

교무 부서는 튜터를 선택 자원으로만 두기보다 수업 활동과 결합하고 사용 목적을 안내하는 방안을 검토할 수 있다(II 장 4 절 라). Hanshaw 와 Sullivan(2025)은 AI 강좌 도우미를 쓰지 않았다는 응답이 많았던 데 대해 실질적 거부보다 효과적이지 못한 소통(안내)이 핵심 문제라고 해석했다. 확대 여부는 대규모 투자나 장기 계약 전에 AI 없이 치른 평가·고차 사고 과제·지연 측정을 둔 학내 효과 측정으로 판단할 수 있다(II 장 4 절 가). Thoeni 와 Fryer(2026)도 대학이 대규모 투자나 장기 계약 전에 AI 챗봇의 교육적 가치를 더 검증해야 한다고 제안했다. AI 에듀테크센터가 운영·실증하는 서비스의 효과는 관계없는 연구자가 분석하거나 분석 계획을 미리 공개하는 방식으로 확인할 만하다.

AI 에듀테크센터와 개인정보 보호·저작권 담당 부서는 강좌 자료를 튜터에 연결하기 전에 이용 허락 범위를 확인하고 대화 로그의 보관·위탁·국외 처리 조건을 계약서에 적으며 화면의 AI 사용 고지를 점검할 수 있다(IV 장 1 절 가·나). 문화체육관광부·한국저작권위원회 안내서는 저작물 일부를 참고 목적으로 입력해도 결과물에 해당 저작물의 중요한 부분이 포함될 수 있어 주의가 필요하다고 안내한다(문화체육관광부·한국저작권위원회, 2025). 개인정보 보호법상 수탁자는 위탁받은 업무 범위를 넘어 개인정보를 이용하거나 제 3 자에게 제공할 수 없다(개인정보 보호법, 2026).

교수자와 교수학습개발센터는 튜터의 한계를 교육과정 목표와 평가에 반영하고 책임 있는 사용을 이끄는 수업 전략을 설계할 수 있다. CS50 Duck 을 살핀 뮌헨응용과학대 연구팀은 책임 있는 생성형 AI 사용을 이끄는 수업 전략과 교수자 관여가 필요하다고 강조했다(Zönnchen et al., 2025).

2026 학년도 2 학기 AI-HiED 교과목(IV 장 2 절 가)의 운영 기록은 AI 에듀테크센터와 교수학습개발센터가 먼저 점검할 수 있다. 이후 교양·전공 강좌에서 튜터를 시범 운영할 때는 수업 활동 결합과 사용 목적 안내, 조교 검수의 공식화, 한국어 강좌 자료 실측, AI 없이 치른 평가·지연 측정을 처음부터 넣고 분석 계획을 미리 공개하는 방안이 있다(부록 4). 교수 업무를 돕는 AI 조교와 학생이 직접 쓰는 튜터는 지원 대상이 달라(IV 장 2 절 가) 점검 항목도 나누어 적용할 만하다.

연구회는 연구책임자 승인과 소프트웨어활용비 집행 가능 여부 확인을 거쳐 두 제품을 수강생·구독자와 같은 조건(Duolingo Max 는 Max 구독과 모바일 앱 환경)에서 실측해 미측정으로 남은

게이트 B3·C1 의 근거를 보완하는 방안을 검토한다. 결과는 채점표 보완판으로 정리해 정기회의에서 공유하고 C5 판정에 필요한 현행 지시 준수 수치와 교수자 개입 경로는 공개 문서로 다시 확인한다.

연구회의 프로그래밍 오류 데이터셋 연구에서 만든 오류 유형 분류는 코딩 튜터의 오개념 교정·이해 점검 과업 문항 세트로 만들어 실측에 쓰는 방안을 검토한다. 교양 AI 교육 프레임워크(연구회 설계안)와는 학습자 주체성·메타인지 영역에 이번 호의 점검 항목을 잇는 방안을 검토하고 실측 절차와 결과는 전공 교과 '공부의 심리'·'AI 와 에듀테크'의 수업 사례로 정리한다. 행동계획(IV 장 1 절 나)에는 학교 현장 교사를 지원하는 과제 60 '교사 지원형 AI 보조교사 도입 및 AI 기반의 교육행정 혁신'이 있다. 이 과제에서 행동계획은 교육부가 2027 년 2 분기까지 AI 튜터를 포함한 AI 기술의 교육적 활용에 대한 윤리·정책 가이드라인을 마련하도록 권고했다(국가인공지능전략위원회, 2026). 연구회는 이번 호의 평가 기준과 기관 도입 점검표(부록 4)를 이 가이드라인 논의의 참고 자료로 정리할 수 있는지 검토한다.

3. 종합

Q1 에 대해, 어떤 설계 요소가 학습 효과를 내는지는 이번 호의 근거로 확정할 수 없고 요소마다 근거 수준이 달랐다. 튜터가 묻는 소크라테스식 질문을 대학 강좌에서 검증한 무작위 실험은 찾지 못했고 학생의 질문을 끌어내는 대화 구조에 단일 회기 실험의 초기 근거가 있었다(II 장 3 절 라). 스캐폴딩을 구조로 넣은 튜터의 개발팀 수행 RCT 두 편은 결과가 일치하지 않았다. 부분별 순차 안내를 넣은 하버드 물리 튜터는 같은 내용의 교실 능동학습보다 큰 효과를 보고했으나 초회 학습의 직후 검사 결과이고 독립 재현은 확인되지 않았다(II 장 1 절 라, II 장 3 절 나). 완전한 해답을 보류한 독일 입문 프로그래밍 튜터(Iris)는 내재 동기를 높였지만 무제한 ChatGPT 와 마찬가지로 통제집단보다 지식 향상의 우위가 없었다(II 장 3 절 나). 두 튜터 모두 여러 요소를 묶은 설계여서 스캐폴딩이나 정답 보류만의 효과는 가를 수 없고 정답 보류형 튜터가 무제한 AI 의 시험 성적 저하를 막은 결과는 K-12 실험에서 나왔다(II 장 3 절 다). 교수자 자료 기반 챗봇은 학기 운영 속 대학 현장 실험(마케팅원론)에서 효과가 없었다(II 장 3 절 다). 오개념 진단은 다단계 문제의 정확도가 낮고 거짓 경보가 많았으며 오류 예제를 더한 규칙 기반 튜터 실험 1 편(대학생 75 명)에서는 향상 폭에 유의한 차이가 없었다(II 장 1 절 다, II 장 3 절 라). 구현 면에서는 정답 보류 지시가 프롬프트만으로 안정되지 않는다는 CS50 개발팀 보고가 있고(2025 년 논문의 GPT-4·GPT-4o 기간 누적 자료에서 응답 22%에 코드 블록) 배치된 서비스는 요청 점검(CS50 Duck)과 응답 사후 점검·재작성(CodeHelp)을 덧대고 있다. 부분별 순차 안내, 설명·풀이 게이팅, 다중 에이전트 점검 같은 구조는 연구용 튜터나 시뮬레이션·벤치마크에서 시험되는 단계이고 교수법 정렬 튜터도 적대적 요청에는 취약했다(II 장 1 절 라). 대표 솔루션에서도 강좌 자료 기반 응답과 요청 점검(CS50 Duck), 전문가가 쓴 시나리오 위의 모델 생성(Duolingo Max)이라는 서로 다른 구조가 문서로 확인됐으나 이 구조가 실제 응답에서 작동하는지는 이번 호에 실측 결과를 싣지 않아 확인되지 않았다(III 장 3 절).

Q2 에 대해, 대학 강좌 실험의 큰 효과는 조건이 붙은 결과였다. 하버드 물리 RCT 의 0.73~1.3 SD 는 이해·적용·분석 수준의 초회 학습에서 구조화된 튜터를 같은 내용의 교실 능동학습과 비교한 값이며 튜터 개발팀이 수행했다(II 장 3 절 나). 학부생 454 명의 무작위 현장 실험에서 시험 성취 효과크기는 d = 0.015 였다(II 장 3 절 다). 챗봇 메타분석에서는 개입 기간의 효과가 엇갈렸다. 한 메타분석은 10 주 미만 개입(ES = 1.179)이 10 주 이상(ES = 0.492)보다 컸다고 했으나(포함 연구 대부분은 ChatGPT 확산 이전 챗봇) 출판 편향을 보정한 다른 메타분석은 1 회성 실험보다 긴 개입의 효과가 컸다고 보고했다(II 장 4 절 가). 고등교육에서 효과가 다른 학교급보다 큰지 작은지는 메타분석마다 결론이 달랐다. 무작위 실험 가운데는 튜터를 실제로 쓴 학생이 14.2%(무료 온라인 강좌)에 그친 경우가 있었다(II 장

4 절 라). 차원별로 보면 큰 정적 결과가 보고된 조건은 구조화된 튜터로 이해·적용·분석 수준 학습 목표를 처음 배운 경우(하버드, 개발팀 수행)다. 입문 과목에서도 무효과(마케팅원론, d = 0.015)와 지식 향상 우위 없음(입문 프로그래밍)이 나와 과목에 따른 경계는 판단할 수 없다. 지식 향상의 우위가 없던 입문 프로그래밍 실험도 90 분의 단기 구조화 실습이어서 인지 수준·기간에 따른 경계 역시 확정할 수 없다. 학습자 특성은 숙련도가 평균 이하인 학습자의 어휘 다양도가 유의하게 오른 하위집단 결과(기업 협력)가 있을 뿐이고 고등교육의 상대적 효과도 메타분석마다 달라 근거가 얇다(Ⅱ 장 3 절 라, Ⅱ 장 4 절 가). 측정 면에서는 AI 없이 치른 평가나 지연 측정으로 효과를 확인한 대학 실험이 적어 사용 중 수행이나 직후 검사의 결과를 학습 효과로 넓혀 읽기 어렵다.

Q3 에 대해, AI 는 반복 질문과 즉시 응답을 맡고 교수자에게 묻기를 망설이는 학생의 질문 창구가 되며 조교는 응답 검수와 심화 지도를, 교원은 교육·지도와 평가 설계를 맡는 분담이 이번 호의 근거와 국내 법제에 비추어 무리가 적은 운영 모델로 보인다(Ⅱ 장 3 절 마, Ⅳ 장 1 절 다). 다만 이 분담의 학습 효과를 검증한 연구는 확인되지 않았고 근거는 개발팀 수행·비무작위·인식 측정에 머문다. 허용 AI 의 범위와 AI 없이 치르는 평가를 정한 강좌 정책(CS50 사례), 수업 활동과 결합한 사용 안내, 도입 전 고지·저작권·위탁·국외 이전 점검이 이 분담을 뒷받침할 수 있다(부록 4). 대학 자체 개발·상용 플랫폼·교수자 설계·개인 구독·인간 튜터 결합 모델은 교수법 통제의 주체와 검수 인력, 데이터 계약, 서비스 지속성에서 차이가 난다(Ⅳ 장 1 절 라). 범용 LLM 을 이미 제공하는 대학은 전용 튜터가 강좌 자료 근거·정답 보류의 구조적 강제·교수자 통제·데이터 계약을 실제로 더 확보하는지를 기준으로 따로 들일지 판단할 수 있다(Ⅳ 장 1 절 라).

이번 호의 근거로 보면 AI 튜터의 교육적 가치는 정답 보류를 구조로 강제하는 설계, 교수자·조교의 검수와 수업 통합, 한국어·데이터·저작권 조건의 점검에 따라 달라질 수 있다. 어느 요소가 학습 효과를 내는지는 아직 무작위 실험으로 확인되지 않았다. 두 대표 솔루션은 자료 기준일까지 공개된 문서로 평가해 정답 보류와 내용 정확성처럼 실측이 필요한 항목은 판정하지 않았다(Ⅰ 장 4 절). 도입 검토 의견은 둘 다 판정 범위가 부족하다는 뜻의 판정 유보(필수 항목 판정 5/13)였다. 게이트 항목 C5 가 두 제품 모두 부분 충족이고 필수 항목 B9(두 제품)와 C4(CS50 Duck)가 미충족이어서 이 판정들이 바뀌지 않는 한 대상 기능을 실측해도 적극 검토에는 이를 수 없다. 이 공백 때문에도 대학은 튜터를 들이기 전에 수강생과 같은 조건에서 정답 보류와 내용 정확성을 직접 확인해야 한다.

다음 호 예고

Vol.04 학습분석 도구(2026 년 9 월분) — 예측에서 개입까지: 학습분석 대시보드·조기경보·학습 기록 저장소(LRS)

부록

부록 1. 참고 자료 목록

가. 국내 자료

개인정보 보호법, 법률 제 21445 호 (2026). https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=283839&chrClsCd=010202&urlMode=lsInfoP&efYd=20260911&ancYnChk=0

고등교육법, 법률 제 21423 호 (2026). https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=283883&chrClsCd=010202&urlMode=lsInfoP&efYd=20260911&ancYnChk=0

공정거래위원회. (2024, 12 월 25 일). 장기 구독권도 환불받을 수 있도록 스픽이지랩스코리아의 이용약관 시정 [보도자료]. 대한민국 정책브리핑. https://www.korea.kr/common/download.do?fileId=197998839&tblKey=GMN

국가인공지능전략위원회. (2026). 대한민국 인공지능 행동계획: 인공지능 기본계획(2026~2028). https://www.aikorea.go.kr/web/board/brdDetail.do?menu_cd=000011&num=328

김가현. (2026, 7 월 29 일). 단국대, 'AI·X 캠퍼스' 본격화… 산학협력 확대. 서울신문. https://www.seoul.co.kr/news/society/education-news/2026/07/29/20260729016002

김소현. (2025, 5 월 26 일). [미래교육혁신 선도대학을 가다] 한림대학교, AI 교육 혁신을 통한 한국형 고등교육 모델 'K-University' 정립. 한국대학신문. https://news.unn.net/news/articleView.html?idxno=579040

김영일. (2026). 규칙·통계 기반과 LLM 기반 맞춤법 검사기의 한국어 오류 처리 패턴 분석 및 교육적 활용 방안. 교양교육연구, 20(1), 211–233. https://doi.org/10.46392/kjge.2026.20.1.211

김의진. (2026, 3 월 4 일). 대학 'AI 교육' 도입 확산… 효율 높인다 vs 강사·조교 인력 축소? 한국대학신문. https://news.unn.net/news/articleView.html?idxno=590453

김혜림, 정향윤. (2026). Human-in-the-Loop(HITL) 기반 대학 글쓰기 교육의 종단적 효과: AI 활용 양상, 튜터링 성취도, 학습 전이를 중심으로. 교양교육연구, 20(2), 95–113. https://doi.org/10.46392/kjge.2026.20.2.95

문화체육관광부, 한국저작권위원회. (2025). 생성형 인공지능 결과물에 의한 저작권 분쟁 예방 안내서 (저작권관계자료 2025-03). https://www.copyright.or.kr/information-materials/publication/research-report/view.do?brdctsno=54252

엘리스. (Retrieved 2026). AI 헬피 [엘리스 LXP 교육자 매뉴얼]. 엘리스 헬프센터. https://help.elice.io/help/docs/elicelxp/for-admin/community/ai-chat

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법, 법률 제 21311 호 (2026). https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=282791&chrClsCd=010202&urlMode=lsInfoP&efYd=20260721&ancYnChk=0

인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령, 대통령령 제 36580 호 (2026). https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=288781&chrClsCd=010202&urlMode=lsInfoP&efYd=20260820&ancYnChk=0

임연서. (2025, 10 월 28 일). "문제 풀이부터 강의내용 요약까지"… 'AI 튜터'로 대학생 학습↑. 한국대학신문. https://news.unn.net/news/articleView.html?idxno=585486

저작권법, 법률 제 21336 호 (2026). https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=283335&chrClsCd=010202&urlMode=lsInfoP&efYd=20260811&ancYnChk=0

저작권법 시행령, 대통령령 제 36569 호 (2026). https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=288567&chrClsCd=010202&urlMode=lsInfoP&efYd=20260811&ancYnChk=0

정재현, 고준보, 신행수. (2025). 원격대학에서 AI 튜터의 도입 가능성과 학습자 인식: 현행 튜터 제도의 개선을 중심으로. 평생학습사회, 21(3), 82–112. https://doi.org/10.26857/jlls.2025.8.21.3.82

조상연, 유연주, 신인섭. (2025). 거대 언어 모델을 활용한 일차방정식 문항의 오개념 기반 오답지 자동 생성. 수학교육, 64(3), 297–313. https://doi.org/10.63311/mathedu.25.6436

한림대학교. (2025, 10 월 16 일). 한림대학교, 'AI 선형대수 튜터' 정식 도입…풀이 과정까지 진단하는 혁신 교육 모델 제시. https://www.hallym.ac.kr/bbs/pr/876/375616/artclView.do

한림대학교. (2026, 6 월 9 일). HAI 에코시스템 서비스, 한림대의 수업 현장으로. https://www.hallym.ac.kr/bbs/fund/889/387809/artclView.do

한림대학교 KELI 사무국. (Retrieved 2026). HAI-MathTutor: AI 기반 초개인화 수학 학습지원 플랫폼 [솔루션 소개]. KELI 플랫폼. https://keliedu.org/keli/solution/collection

황영아, 신종호. (2025). AI 기반 말하기 앱의 효과성 분석과 대학 교양 영어에서의 활용 가능성 탐색: 비교과 프로그램 운영 결과를 중심으로. 교양교육연구, 19(3), 53–72. https://doi.org/10.46392/kjge.2025.19.3.53

나. 국외 자료

1EdTech Consortium. (Retrieved 2026). Duolingo [Product certification listing]. TrustEd Apps Directory. https://site.imsglobal.org/certifications/duolingo/duolingo

Alemdağ, E. (2025). The effect of chatbots on learning: A meta-analysis of empirical research. Journal of Research on Technology in Education, 57(2), 459–481. https://doi.org/10.1080/15391523.2023.2255698

Anthropic. (2025a, July 9). Advancing Claude for Education. https://www.anthropic.com/news/advancing-claude-for-education

Anthropic. (2025b, October 23). Expanding our use of Google Cloud TPUs and services. https://www.anthropic.com/news/expanding-our-use-of-google-cloud-tpus-and-services

Bassner, P., Lenk-Ostendorf, B., Beinstingel, R., Wasner, T., & Krusche, S. (2026). Less stress, better scores, same learning: The dissociation of performance and learning in AI-supported programming education. Computers and Education: Artificial Intelligence, 10, Article 100537. https://doi.org/10.1016/j.caeai.2025.100537

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122

Chegg. (2026). Annual report on Form 10-K for the fiscal year ended December 31, 2025. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1364954/000136495426000021/chgg-20251231.htm

Cheon, G., Choi, Y., Lee, D., & Baek, J. (2025). Generative AI agents in language learning: A randomized field experiment. In Proceedings of the 58th Hawaii International Conference on System Sciences (pp. 4949–4958). Hawaii International Conference on System Sciences. https://doi.org/10.24251/hicss.2025.596

Chu, Z., Wang, S., Xie, J., Zhu, T., Yan, Y., Ye, J., Zhong, A., Hu, X., Liang, J., Yu, P. S., & Wen, Q. (2025). LLM agents for education: Advances and applications. In Findings of the Association for Computational Linguistics: EMNLP 2025 (pp. 13782–13810). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-emnlp.743

Competition and Markets Authority. (2024). Alphabet Inc.'s partnership with Anthropic PBC: Decision on relevant merger situation (ME/7108/24). https://assets.publishing.service.gov.uk/media/676959bae6ff7c8a1fde9d33/Full_text_decision__.pdf

CS50. (Retrieved 2026). CS50.ai. CS50 Docs. https://cs50.readthedocs.io/cs50.ai/

CS50. (2026). Syllabus [Harvard College, Fall 2026]. https://cs50.harvard.edu/college/2026/fall/syllabus/

Daheim, N., Macina, J., Kapur, M., Gurevych, I., & Sachan, M. (2024). Stepwise verification and remediation of student reasoning errors with large language model tutors. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 8386–8411). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.478

Dahlkemper, M. N., Lahme, S. Z., & Klein, P. (2023). How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT. Physical Review Physics Education Research, 19(1), Article 010142. https://doi.org/10.1103/physrevphyseducres.19.010142

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. Computers & Education, 210, Article 104967. https://doi.org/10.1016/j.compedu.2023.104967

De Simone, M., Tiberti, F., Barron Rodriguez, M., Manolio, F., Mosuro, W., & Dikoru, E. J. (2025). From chalkboards to chatbots: Evaluating the impact of generative AI on learning outcomes in Nigeria (Policy Research Working Paper No. 11125). World Bank. https://doi.org/10.1596/1813-9450-11125

del Rio-Chanona, R. M., Laurentsyeva, N., & Wachs, J. (2024). Large language models reduce public knowledge sharing on online Q&A platforms. PNAS Nexus, 3(9), Article pgae400. https://doi.org/10.1093/pnasnexus/pgae400

Department for Science, Innovation and Technology & Department for Education. (2026, April 16). Edtech and AI companies invited to help build safe AI tutoring tools for disadvantaged pupils [Press release]. GOV.UK. https://www.gov.uk/government/news/edtech-and-ai-companies-invited-to-help-build-safe-ai-tutoring-tools-for-disadvantaged-pupils

Di Lonardo Burr, S. M., Douglas, H., Vorobeva, M., & Muldner, K. (2020). Refuting misconceptions: Computer tutors for fraction arithmetic. Journal of Numerical Cognition, 6(3), 355–377. https://doi.org/10.5964/jnc.v6i3.310

Diel, A., Torous, J., Cuijpers, P., Kleesiek, J., Nensa, F., Weber, N., Faust, F., Lalgi, T. J., Mellis, F. S., Teufel, M., & Bäuerle, A. (2026). A scoping review on the mental health harms of LLM-based chatbots. npj Digital Medicine, 9(1), Article 644. https://doi.org/10.1038/s41746-026-03054-x

Dinucu-Jianu, D., Macina, J., Daheim, N., Hakimi, I., Gurevych, I., & Sachan, M. (2025). From problem-solving to teaching problem-solving: Aligning LLMs with pedagogy using reinforcement learning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 272–292). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.15

Duolingo. (Retrieved 2026-a). 듀오링고 맥스란 무엇인가요? Duolingo 도움말 센터. https://www.duolingo.com/help/what-is-duolingo-max

Duolingo. (Retrieved 2026-b). 슈퍼 듀오링고란 무엇이며 구독은 어떻게 하나요? Duolingo 도움말 센터. https://www.duolingo.com/help/super-duolingo

Duolingo. (2023, March 14). Introducing Duolingo Max, a learning experience powered by GPT-4. Duolingo Blog. https://blog.duolingo.com/duolingo-max/

Duolingo. (2025, October 27). 서비스 약관. https://www.duolingo.com/terms

Duolingo. (2026a). Q4/FY 2025 shareholder letter [Exhibit 99.2 to Form 8-K]. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1562088/000162828026012246/q4fy25duolingo12-31x25shar.htm

Duolingo. (2026b). Annual report on Form 10-K for the fiscal year ended December 31, 2025. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1562088/000162828026012494/duol-20251231.htm

Duolingo. (2026c, May 26). 개인정보처리방침. https://www.duolingo.com/privacy

Duolingo. (2026d). Q2 2026 shareholder letter [Exhibit 99.2 to Form 8-K]. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1562088/000162828026053299/q2fy26duolingo6-30x26share.htm

Duolingo. (2026e). Quarterly report on Form 10-Q for the quarterly period ended June 30, 2026. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/1562088/000162828026053603/duol-20260630.htm

Duolingo. (2026f, August 24). Important information regarding Duolingo for Schools. Duolingo for Schools Help Center. https://duolingoschools.zendesk.com/hc/en-us/articles/44824545841805-Important-information-regarding-Duolingo-for-Schools

Duolingo. (2026g, August 30). Is Duolingo compliant with my school or district's privacy policies? Duolingo for Schools Help Center. https://duolingoschools.zendesk.com/hc/en-us/articles/6830532051085-Is-Duolingo-compliant-with-my-school-or-district-s-privacy-policies

European Commission. (2025). Commission guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act) (C(2025) 5052 final). https://ec.europa.eu/newsroom/dae/redirection/document/112367

Gedera, D., & Griffiths, C. (2026). Reshaping business education: An activity theory analysis of AI teaching assistants. Research and Practice in Technology Enhanced Learning, 22, Article 004. https://doi.org/10.58459/rptel.2027.22004

Grand View Research. (2025). AI tutors market size, share & trends analysis report by application (subject-specific tutoring, test preparation), by technology (natural language processing, machine learning and predictive analytics), by end-use, by region, and segment forecasts, 2025–2030 (Report ID GVR-4-68040-495-2). https://web.archive.org/web/20250730030807/https://www.grandviewresearch.com/industry-analysis/ai-tutors-market-report

Hanshaw, G., & Sullivan, C. (2025). Exploring barriers to AI course assistant adoption: A mixed-methods study on student non-utilization. Discover Artificial Intelligence, 5(1), Article 178. https://doi.org/10.1007/s44163-025-00312-x

Imran, M., & Bulathwela, S. (2026). Catching the correct answer trap: Characterising AI tutor blind spots when analysing student reasoning. In E. G. Blanchard, G. Chen, M. Chi, & S. Isotani (Eds.), Artificial intelligence in education: 27th International Conference, AIED 2026, Seoul, South Korea, June 27–July 3, 2026, proceedings, part V (pp. 383–392). Springer. https://doi.org/10.1007/978-3-032-29770-9_42

Kang, B. O., Jeon, H.-B., & Lee, Y. K. (2024). AI-based language tutoring systems with end-to-end automatic speech recognition and proficiency evaluation. ETRI Journal, 46(1), 48–58. https://doi.org/10.4218/etrij.2023-0322

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Khan Academy. (2026, April 21). Learning in the open: What AI is (and isn't) changing. Khan Academy Blog. https://blog.khanacademy.org/learning-in-the-open-what-ai-is-and-isnt-changing/

Kim, Y., Shin, S., & Yoo, S.-H. (2025). Performance of large language models in non-English medical ethics-related multiple choice questions: Comparison of ChatGPT performance across versions and languages. BMC Medical Ethics, 26(1), Article 168. https://doi.org/10.1186/s12910-025-01316-z

Kittredge, A. K., Hopman, E. W. M., Reuveni, B., Dionne, D., Freeman, C., & Jiang, X. (2025). Mobile language app learners' self-efficacy increases after using generative AI. Frontiers in Education, 10, Article 1499497. https://doi.org/10.3389/feduc.2025.1499497

Kittredge, A., Lee, C., & Jiang, X. (2025). Video Call improves Japanese English learners' speaking skills (Duolingo Research Report DRR-25-06). Duolingo. https://web.archive.org/web/20260415005444/https://duolingo-papers.s3.amazonaws.com/reports/Duolingo_whitepaper_language_video_call_improves_speaking_2025.pdf?ref=blog.duolingo.com

Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. Review of Educational Research, 86(1), 42–78. https://doi.org/10.3102/0034654315581420

Kweon, S., Nam, S., Lim, H., Hong, H., & Choi, E. (2025). A large-scale real-world evaluation of an LLM-based virtual teaching assistant. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track) (pp. 850–864). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-industry.60

Laun, M., & Wolff, F. (2025). Chatbots in education: Hype or help? A meta-analysis. Learning and Individual Differences, 119, Article 102646. https://doi.org/10.1016/j.lindif.2025.102646

Lee, H., & Kwon, H. (2026). Instructor-designed AI tutors in university foreign language education: A mixed-methods study of learner motivation and reflective learning experience based on self-determination theory. Trends in Higher Education, 5(3), Article 78. https://doi.org/10.3390/higheredu5030078

Lee, J., Youn, G., Lee, S., Im, C., Kim, J., & Yoo, C. (2026). Let LLM tutors ask first: Proactive LLM-based tutoring at scale in a 1,500-student online classroom. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track) (pp. 1541–1554). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-industry.107

Li, Z., Wang, Z., Wang, W., Hung, K., Xie, H., & Wang, F. L. (2025). Retrieval-augmented generation for educational application: A systematic survey. Computers and Education: Artificial Intelligence, 8, Article 100417. https://doi.org/10.1016/j.caeai.2025.100417

Liffiton, M., Sheese, B. E., Savelka, J., & Denny, P. (2023). CodeHelp: Using large language models with guardrails for scalable support in programming classes. In Proceedings of the 23rd Koli Calling International Conference on Computing Education Research (pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3631802.3631830

Lindgren, R., Kakar, S., Maiti, P., Taneja, K., & Goel, A. K. (2024). Does Jill Watson increase teaching presence? In Proceedings of the Eleventh ACM Conference on Learning @ Scale (pp. 269–273). Association for Computing Machinery. https://doi.org/10.1145/3657604.3664679

Liu, C. H., Zhao, C., Lou, F., Zhang, W., Menor, A., Chow, A., & Yip, T. (2026). Loneliness and generative AI use for companionship among college students. Computers in Human Behavior: Artificial Humans, 9, Article 100348. https://doi.org/10.1016/j.chbah.2026.100348

Liu, R., Zenke, C., Liu, C., Holmes, A., Thornton, P., & Malan, D. J. (2024). Teaching CS50 with AI: Leveraging generative artificial intelligence in computer science education. In Proceedings of the 55th ACM

Technical Symposium on Computer Science Education V. 1 (pp. 750–756). Association for Computing Machinery. https://doi.org/10.1145/3626252.3630938

Liu, R., Zhao, J., Xu, B., Perez, C., Zhukovets, Y., & Malan, D. J. (2025). Improving AI in CS50: Leveraging human feedback for better learning. In Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1 (pp. 715–721). Association for Computing Machinery. https://doi.org/10.1145/3641554.3701945

Lunden, I. (2024, December 10). OpenAI-backed Speak raises $78M at $1B valuation to help users learn languages by talking out loud. TechCrunch. https://techcrunch.com/2024/12/10/openai-backed-speak-raises-78m-at-1b-valuation-to-help-users-learn-languages-by-talking-out-loud/

Lyu, W., Wang, Y., Chung, T., Sun, Y., & Zhang, Y. (2024). Evaluating the effectiveness of LLMs in introductory computer science education: A semester-long field study. In Proceedings of the Eleventh ACM Conference on Learning @ Scale (pp. 63–74). Association for Computing Machinery. https://doi.org/10.1145/3657604.3662036

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901–918. https://doi.org/10.1037/a0037123

Macina, J., Daheim, N., Hakimi, I., Kapur, M., Gurevych, I., & Sachan, M. (2025). MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 204–221). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.11

Mas Castillo, L. (2026, January 1). Explain My Answer is now free for all learners! Duolingo Blog. https://blog.duolingo.com/explain-my-answer-now-free/

Maurya, K. K., Srivatsa, K. V. A., Petukhova, K., & Kochmar, E. (2025). Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 1234–1251). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.57

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. UNESCO. https://doi.org/10.54675/ewzm9535

Microsoft. (2026, May 18). Data, privacy, and security for Models sold by Azure in Microsoft Foundry. Microsoft Learn. https://learn.microsoft.com/en-us/azure/foundry/responsible-ai/openai/data-privacy

Microsoft Corporate Blogs. (2025, November 18). Microsoft, NVIDIA and Anthropic announce strategic partnerships. The Official Microsoft Blog. https://blogs.microsoft.com/blog/2025/11/18/microsoft-nvidia-and-anthropic-announce-strategic-partnerships/

Ng, B. Y., Li, J., Tong, X., Ye, K., Yenne, G., Chandrasekaran, V., & Li, J. (2026). Understanding how university guidelines address privacy and security issues of generative AI in academic settings. Proceedings on Privacy Enhancing Technologies, 2026(4), 1055–1071. https://doi.org/10.56553/popets-2026-0159

Nie, A., Chandak, Y., Suzara, M., Malik, A., Woodrow, J., Peng, M., Sahami, M., Brunskill, E., & Piech, C. (2025). The GPT surprise: Offering large language model chat in a massive coding class reduced engagement but may increase adopters' exam performances. In Proceedings of the Twelfth ACM Conference on Learning @ Scale (pp. 376–380). Association for Computing Machinery. https://doi.org/10.1145/3698205.3733960

OECD. (2026a). OECD digital education outlook 2026: Exploring effective uses of generative AI in education. OECD Publishing. https://doi.org/10.1787/062a7394-en

OECD. (2026b). Policies supporting responsible and systematic GenAI adoption in higher education (OECD Education Spotlights No. 23). OECD Publishing. https://doi.org/10.1787/c4e5621f-en

Oreopoulos, P., & Low, N. (2026). One click away: AI tutoring with Khanmigo in a two-year school experiment (NBER Working Paper No. 35620). National Bureau of Economic Research. https://doi.org/10.3386/w35620

Ouyang, Z., Jiang, Y., & Liu, H. (2024). The effects of Duolingo, an AI-integrated technology, on EFL learners' willingness to communicate and engagement in online classes. The International Review of Research in Open and Distributed Learning, 25(3), 97–115. https://doi.org/10.19173/irrodl.v25i3.7677

Pardos, Z. A., & Bhandari, S. (2024). ChatGPT-generated help produces learning gains equivalent to human tutor-authored help on mathematics skills. PLOS ONE, 19(5), Article e0304013. https://doi.org/10.1371/journal.pone.0304013

Park, S. (2026). Intent vs. surface: Recovering acoustic realization from modern ASR for pronunciation training. In Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026) (pp. 336–346). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.bea-1.23

Park, S., & Kim, G. (2025). Evaluating multimodal generative AI with Korean educational standards. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers) (pp. 671–688). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-short.56

Reddig, J. M., Arora, A., & MacLellan, C. J. (2025). Generating in-context, personalized feedback for intelligent tutors with large language models. International Journal of Artificial Intelligence in Education, 35(6), 3459–3500. https://doi.org/10.1007/s40593-025-00505-6

Retraction: Gamified mobile learning: An escape room chatbot to enhance medical terminology acquisition. (2026). Health Education Journal, 85(6), 741. https://doi.org/10.1177/00178969261481011

Schroeder, N. L., & Kucera, A. C. (2022). Refutation text facilitates learning: A meta-analysis of between-subjects experiments. Educational Psychology Review, 34(2), 957–987. https://doi.org/10.1007/s10648-021-09656-z

Sok, S., & Shin, H. W. (2025). Do interactions with ChatGPT influence L2 learners' oral speaking ability, summarization ability, and perceptions of generative AI tasks? TESOL Quarterly, 59(S1). https://doi.org/10.1002/tesq.70001

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. Journal of Educational Psychology, 106(2), 331–347. https://doi.org/10.1037/a0034752

Taneja, K., Maiti, P., Kakar, S., Guruprasad, P., Rao, S., & Goel, A. K. (2024). Jill Watson: A virtual teaching assistant powered by ChatGPT. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), Artificial intelligence in education: 25th International Conference, AIED 2024, Recife, Brazil, July 8–12, 2024, proceedings, part I (pp. 324–337). Springer. https://doi.org/10.1007/978-3-031-64302-6_23

Thoeni, A., & Fryer, L. K. (2026). AI chatbots in higher education: Comparing expectations to evidence. Computers in Human Behavior Reports, 22, Article 101061. https://doi.org/10.1016/j.chbr.2026.101061

Turner, C. (2026, June 17). 'Limited' evidence for AI tutoring tools, government admits. Tes. https://www.tes.com/magazine/news/general/limited-evidence-ai-tutoring-tools-government-admits

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Wang, F., Cheung, A. C. K., Neitzel, A. J., & Chai, C. S. (2025). Does chatting with chatbots improve language learning performance? A meta-analysis of chatbot-assisted language learning. Review of Educational Research, 95(4), 623–660. https://doi.org/10.3102/00346543241255621

Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. British Journal of Educational Technology, 55(1), 10–33. https://doi.org/10.1111/bjet.13334

Yang, J. W., Choe, K., Lim, J., & Park, J. (2025). Asking questions matters: Comparing the effect of learning through dialogue with peers, ChatGPT, and intelligent tutoring system. Interactive Learning Environments, 33(10), 5870–5886. https://doi.org/10.1080/10494820.2025.2488984

Yasir, T., Li, W., Gilson, S., Tithi, S. D., Tian, X., & Barnes, T. (2026). Confirming correct, missing the rest: LLM tutoring agents struggle where feedback matters most. In Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026) (pp. 819–840). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.bea-1.56

Zhao, J., Knežević, M., & Käser, T. (2026). Evaluating answer leakage robustness of LLM tutors against adversarial student attacks. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 30588–30617). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.1412

Zhao, S., Yu, K., Yuan, Y., He, P., & Wen, H. (2026). SHAPE: Unifying safety, helpfulness and pedagogy for educational LLMs. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 11537–11553). Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.acl-long.529

Zönnchen, B., Hobelsberger, M., Socher, G., Thurner, V., & Ottinger, S. (2025). Exploring the role of large language models as artificial tutors. In 2025 IEEE Global Engineering Education Conference (EDUCON) (pp. 1–10). IEEE. https://doi.org/10.1109/educon62633.2025.11016571

부록 2. 평가척도 채점표

두 대표 솔루션의 항목별 판정과 근거는 다음과 같다. 판정은 자료 기준일까지 공개된 문서와 연구에 근거한다.

대상 솔루션 CS50 Duck (Harvard CS50) **평가자** 연구진(문서 판정, 채점 초안 작성에 생성형 AI 보조) **접근 방식** 자료 기준일까지 공개된 제공사 문서·정책(CS50 공식 문서·강좌 실라버스 등)과 연구(CS50 팀 논문·독립 연구)로 판정했다(실사용 진단 결과는 이번 호에 싣지 않음). 실사용 측정이 필요한 14 개 항목은 판정하지 않았다.

ID	항목	측정	가중	판정	근거·출처
A. 사용성 (가중치 30%)					
A1	과업 완료 성과	AUTO	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A2	효율성·반응성	AUTO	권장	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).

ID	항목	측정	가중	판정	근거·출처
A3	시스템 상태 가시성	HYBRID	권장	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A4	오류 예방·복구	HYBRID	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A5	일관성·표준 준수	HYBRID	보조	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A6	접근성 — 자동 검출 영역	AUTO	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A7	접근성 — 수동 검증 영역	HYBRID	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A8	주관적 만족도(SUS 등)	DOC	권장	◐	제공사와 독립된 뮌헨응용과학대 연구팀이 컴퓨팅 사고 강좌에서 학생 인식·참여·실제 활용을 살핀 혼합연구의 설문·과제 자료에서 학생들은 CS50 Duck 의 비판단적 피드백을 높이 평가했다. SUS 등 표준 사용성 점수는 공개되지 않아 부분 충족으로 판정했다. 참고: CS50 팀이 수행한 설문(2023 년 가을 캠퍼스 약 500 명, 비익명, 자체 보고)에서는 학생 47%가 '매우 도움됨', 26%가 '도움됨'으로 답했다. (출처: Zönnchen et al., 2025; R. Liu et al., 2024)
B. 교육적 특성 (가중치 40%)					
B1	학습목표 정렬	DOC	필수	◐	2024 년 논문에 따르면 CS50.ai 는 GPT-4 요청마다 강좌별 규칙을 담은 시스템 프롬프트를 보내고 강의 관련 질문에는 CS50 강의 자막 기반 검색 증강 생성(RAG)으로 답한다. 2025 년 논문은 조교가 정기적으로 검토하는 모델 평가 체계로 교수 목표와의 정렬을 유지한다고 밝힌다. 그러나 같은 2025 년 논문(자체 보고)은 모델이 복잡한 시스템 지시를 지키지 못하는 지시 희석과 GPT-4o 기간 응답의 코드 블록 비율(메시지 기준 25%·대화 기준 56%)을 보고했다. 같은 논문은 소규모 고품질 데이터셋 미세조정으로 다회차 대화 평가에서 교수 목표와의 정렬이 크게 개선됐다고 보고했다. 공식 문서가 표방하는 것은 학생을 답으로 이끄는 튜터라는 설계 지향이고 CS50 Duck 의 명시적 학습 목표와 CS50 Duck 에 연결된 평가는 문서로 확인되지 않았다(2026 년 가을 실라버스는 퀴즈를 강의실 대면 응시로 안내). 강좌 규칙·강의 자막에 맞춘 설계는 확인되나 부분 충족으로 판정했다. (출처: R. Liu et al., 2024; R. Liu et al., 2025; CS50, Retrieved 2026; CS50, 2026)

ID	항목	측정	가중	판정	근거·출처
B2	콘텐츠 품질·정확성	HYBRID	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
B3	피드백·적응	HYBRID	필수*	미측정	(게이트 항목) 실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음). 참고: 공식 문서는 CS50 Duck 이 답을 곧바로 주기보다 학생을 답으로 이끄는 튜터처럼 설계되었다고 밝힌다. CS50 팀의 분석(자체 보고)에서는 GPT-4o 기간 응답의 코드 블록 비율이 메시지 기준 25%·대화 기준 56%였다. (출처: CS50, Retrieved 2026; R. Liu et al., 2025)
B4	동기·참여	DOC	권장	◐	제공사와 독립된 뮌헨응용과학대 연구팀의 연구에서 학생들은 CS50 Duck 의 비판단적 피드백을 높이 평가했으나 어려움을 겪는 학생은 과의존해 핵심 학습 과정을 건너뛸 수 있다고 보고되었다. CS50 팀도 잦은 코드 블록 생성이 학생의 CS50 Duck 의존을 키우고 문제 해결 과정 참여를 막을 수 있다고 우려했다(자체 보고). 참여를 이끄는 요인과 과의존 위험이 함께 보고되어 부분 충족으로 판정했다. (출처: Zönnchen et al., 2025; R. Liu et al., 2025)
B5	교수설계·인지부하	HYBRID	권장	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
B6	교수 통합·교수자 지원	DOC	권장	◐	2026 년 가을 실라버스는 CS50 Duck 을 과제의 허용 AI 로 지정하고 섹션을 예약제 오피스아워로 보완하며 관리자가 조교에게 모델 응답 비교 평가를 배정하는 체계도 있다. 공식 문서는 CS50.ai 를 학생·교사용 ChatGPT 변형으로 소개하지만 다른 기관 교수자가 CS50 Duck 을 자기 강좌에 맞게 설정하는 기능이나 교사용 자료·차별화 지원은 문서로 확인되지 않았다. 강좌 안 통합은 확인되나 외부 교수자 지원이 부족해 부분 충족으로 판정했다. (출처: CS50, 2026; R. Liu et al., 2025; CS50, Retrieved 2026)
B7	상호작용 사용성·접근성(교육 맥락)	HYBRID	보조	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음). 참고: 공식 문서는 CS50 Duck 을 cs50.ai 웹과 CS50 용 Visual Studio Code(cs50.dev)에서 제공한다고 밝힌다. (출처: CS50, Retrieved 2026)
B8	표준 준수·상호운용	DOC	보조	○	공식 문서가 밝히는 제공 경로는 cs50.ai 웹과 CS50 용 Visual Studio Code(cs50.dev)다. LTI 등 교육 데이터·콘텐츠 상호운용 표준의 지원이나 인증은 문서로 확인되지 않아 미충족으로 판정했다. (출처: CS50, Retrieved 2026)

ID	항목	측정	가중	판정	근거·출처
B9	효과성 증거 등급	DOC	필수	○	제공사 독립·고등교육 대상·제품 단위 조건을 갖춘 연구는 뮌헨응용과학대 컴퓨팅 사고 강좌에서 학생 인식·참여·활용을 살핀 혼합연구 1 편이다. 이 연구는 초록 기준 통제집단과 효과크기가 없어 학습 성과 근거가 되지 못하므로 독립 효과 근거 없음으로 판정했다. 참고: CS50 팀이 수행한 논문의 학생 설문과 조교의 응답 비교 평가는 자체 보고라 반영하지 않았다. (출처: Zönnchen et al., 2025; R. Liu et al., 2024; R. Liu et al., 2025)
B10	효과크기 정량	DOC	권장	○	제공사 독립·고등교육 대상 연구(뮌헨응용과학대 혼합연구)는 초록 기준 학습 성과 효과크기를 보고하지 않아 미충족으로 판정했다. 참고: CS50 팀이 수행한 논문 2 편도 학습 성과 효과크기는 보고하지 않았다. (출처: Zönnchen et al., 2025; R. Liu et al., 2024; R. Liu et al., 2025)
C. AI 신뢰성·윤리·거버넌스 (가중치 30%)					
C1	타당성·신뢰성과 환각 관리	HYBRID	필수*	미측정	(게이트 항목) 실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음). 참고: CS50 팀은 강의 자막 기반 RAG 로 강의 관련 답변의 환각 가능성을 줄였다고 밝힌다. 같은 팀의 보고에서 2023 년 가을 Ed 게시판 답변 180 건 중 직원 승인은 70 건(39%, 저자는 Ed 이용 감소로 실제 정확도를 과소평가한 값으로 봄) 이었고 모델이 복잡한 시스템 지시를 지키지 못하는 지시 희석 문제도 확인되었다. (출처: R. Liu et al., 2024; R. Liu et al., 2025)
C2	투명성·설명가능성	DOC	필수	◐	공식 문서는 CS50.ai 를 학생·교사용 ChatGPT 변형으로 소개하고 시스템 프롬프트·강의 자막 RAG·Azure 호스팅 GPT-4o 중계 구조도 논문으로 공개되었다. 그러나 공식 문서는 기반 모델을 ChatGPT 로만 불러 현행 모델명은 확인되지 않고 응답의 근거를 학습자에게 보여주는 방식도 문서로 확인되지 않아 부분 충족으로 판정했다. (출처: CS50, Retrieved 2026; R. Liu et al., 2024; R. Liu et al., 2025)
C3	공정성·편향 관리	HYBRID	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).

ID	항목	측정	가중	판정	근거·출처
C4	프라이버시·데이터 거버넌스	DOC	필수	○	공식 문서·실라버스에는 대화 기록의 보관 기간·이용 목적·삭제·연령 조항이 없으며 CS50 팀이 학생 대화 메시지 1,000 만 건을 분석한 이용에 관한 고지도 확인되지 않았다. CS50.ai 가 학생 메시지를 중계하는 Azure 의 경우 Microsoft 는 프롬프트·응답을 OpenAI 에 제공하지 않고 허락 없이 기반 모델 학습에 쓰지 않는다고 밝히지만 처리자 약정이라 CS50 의 고지를 대신하지 못해 미충족으로 판정했다. (출처: CS50, Retrieved 2026; CS50, 2026; R. Liu et al., 2025; Microsoft, 2026)
C5	인간 통제·교수자 역할	DOC	필수*	◐	(게이트 항목) 관리자가 배정한 평가 과제로 조교가 같은 학생 메시지에 대한 모델 응답을 비교·선택하는 정기 검토 체계가 있다. 2026 년 가을 실라버스는 과제에서 CS50 자체 AI 만 허용하고 섹션을 예약제 오피스아워로 보완하며 퀴즈를 강의실 대면으로 치르도록 안내한다. 그러나 CS50 팀은 모델이 복잡한 시스템 지시를 지키지 못하는 지시 희석과 GPT-4o 기간의 코드 블록 비율(메시지 기준 25%·대화 기준 56%, 자체 보고)을 보고했고 OpenAI 의 기반 모델 갱신이 튜터 행동을 예기치 않게 바꿀 수 있다고 밝혔다. AI 응답에 대한 사람의 관여는 모델 비교 평가로만 확인되어 교수진의 지시가 응답에 일관되게 반영된다고 보기 어려워 부분 충족으로 판정했다. (출처: R. Liu et al., 2025; CS50, 2026)
C6	안전·견고성·보안	HYBRID	권장	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
C7	책무성·거버넌스·정책 정합	DOC	권장	◐	공식 문서는 CS50 Duck 을 CS50 자체 AI 소프트웨어로 안내해 운영 주체가 분명하고 CS50 팀은 조교가 정기적으로 검토하는 모델 평가 체계를 두며 OpenAI 의 기반 모델 갱신이 튜터 행동을 예기치 않게 바꿀 수 있다고 밝혔다. 그러나 AI 응답 오류에 대한 이의 제기 절차·감사 기록 운용과 한국 이용자 대상 고지 등 국내 정책 대응은 문서로 확인되지 않아 부분 충족으로 판정했다. (출처: CS50, Retrieved 2026; R. Liu et al., 2025)
C8	AI 리터러시·연령 적합성	DOC	보조	○	공식 문서와 실라버스는 CS50 자체 도구 외의 AI 로 답이나 코드를 제안·완성받는 것을 원칙적으로 허용하지 않는다고 정해 학습자가 쓸 수 있는 AI 의 범위를 안내한다. 그러나 이는 학업 정직성 규정이어서 AI 리터러시 지원으로 보기 어렵고 cs50.ai 의 독립 사용 연령 기준과 학습자용 AI 리터러시 자료가 모두 문서로 확인되지 않아 미충족으로 판정했다. (출처: CS50, Retrieved 2026; CS50, 2026)

ID	항목	측정	가중	판정	근거·출처
축 점수 A 50.0 · B 26.9 · C 33.3 → 종합 35.8 | 고부담 게이트 해당 없음 | 도입 검토 의견 판정 유보					

자료 기준일까지 공개된 문서로 26 개 항목 중 12 개(필수 13 개 중 5 개)를 판정했다. 실사용 측정이 필요한 14 개 항목(게이트 B3·C1 포함)은 미측정으로 점수에서 뺐다. 필수 항목 13 개 가운데 판정한 항목이 5 개로 경계값 적용 최소 범위(10 개)에 미치지 못해 종합 35.8 점(축 A 는 A8 한 항목의 값)과 관계없이 도입 검토 의견은 '판정 유보'다. 게이트 항목 C5 가 ◐이고 필수 항목 B9·C4 가 ○여서 이 판정들이 바뀌지 않는 한 미측정 항목을 측정해도 적극 검토에는 이를 수 없다. 공통 항목 기준 29.6 점은 두 제품이 모두 판정한 11 개 항목(축 B·C)만으로 점수가 있는 축의 가중치(40·30)로 다시 계산한 비교용 점수다. 축 A 가 이 제품에만 있으므로 두 제품의 종합 점수끼리는 직접 비교하지 않는다. 축 점수는 소수 첫째 자리까지 반올림하고 종합 점수는 반올림한 축 점수로 계산했다.

대상 솔루션 Duolingo Max (Duolingo) **평가자** 연구진(문서 판정, 채점 초안 작성에 생성형 AI 보조) **접근 방식** 자료 기준일까지 공개된 제공사 문서·정책(공식 블로그·도움말·개인정보처리방침·서비스 약관·연차보고서 등)과 연구(제공사 연구 보고서·독립 연구)로 판정했다(실사용 진단 결과는 이번 호에 싣지 않음). 실사용 측정이 필요한 14 개 항목은 판정하지 않았다.

ID	항목	측정	가중	판정	근거·출처
A. 사용성 (가중치 30%)					
A1	과업 완료 성과	AUTO	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A2	효율성·반응성	AUTO	권장	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A3	시스템 상태 가시성	HYBRID	권장	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A4	오류 예방·복구	HYBRID	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A5	일관성·표준 준수	HYBRID	보조	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A6	접근성 — 자동 검출 영역	AUTO	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
A7	접근성 — 수동 검증 영역	HYBRID	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).

ID	항목	측정	가중	판정	근거·출처
A8	주관적 만족도(SUS 등)	DOC	권장	미측정	Duolingo Max 기능을 대상으로 한 SUS 등 공개 사용성·만족도 조사와 제공사와 독립된 고등교육 대상 만족도 연구를 찾지 못했고 소규모 사용자 평정도 하지 않아 판정하지 않았다. 참고: Duolingo 소속 연구팀의 Video Call 무작위 통제 실험 보고서(B1.1 영어 과정 일본어 화자 567 명, 자체 보고, 동료심사 전)는 참여자들이 사후 설문에서 Video Call 기능을 중간 정도로 긍정적으로 인식했다고 밝혔다. 같은 보고서에서 Video Call 조건의 탈락률(20.1%)은 통제 조건(7.6%)보다 높았다. (출처: Kittredge, Lee, & Jiang, 2025)
B. 교육적 특성 (가중치 40%)					
B1	학습목표 정렬	DOC	필수	◐	Roleplay 시나리오와 첫 메시지는 Duolingo 전문가가 써서 학습자의 과정 위치에 맞추고 모델에 대화 방향을 지시하며 Explain My Answer 는 레슨 안에서 학습자의 오류에 맞춘 피드백을 준다. 제공사가 Duolingo Max 의 핵심 기능으로 소개하는 Video Call 이 과정의 학습 목표와 어떻게 정렬되는지는 문서로 확인되지 않아 부분 충족으로 판정했다. (출처: Duolingo, 2023; Mas Castillo, 2026; Duolingo, 2026a)
B2	콘텐츠 품질·정확성	HYBRID	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
B3	피드백·적응	HYBRID	필수*	미측정	(게이트 항목) 실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음). 참고: Duolingo 는 Explain My Answer 가 학습자의 오류와 규칙의 어려운 점, 모국어 배경을 고려한 개인화 피드백을 만들고 피드백을 받을지 학습자가 고를 수 있다고 설명한다. (출처: Mas Castillo, 2026)
B4	동기·참여	DOC	권장	◐	Duolingo 는 AI 생성 내용이 재미있고 격려하는 어조인지 검토하고 Video Call 캐릭터가 이전 통화 내용을 기억한다고 밝힌다. 제공사와 독립된 중국 대학 1 학년 준실험에서 일반 Duolingo 앱 사용의 학습 참여 총점 효과크기는 Cohen's d = 2.03 이었으나 Duolingo Max 기능을 다룬 연구는 아니다. 자체 보고(동료심사 전)의 Video Call 실험에서는 Video Call 조건의 탈락률(20.1%)이 통제 조건(7.6%)보다 높았고 한 달 사용으로 말하기 자신감이 높아지지 않아 부분 충족으로 판정했다. (출처: Duolingo, 2023; Ouyang et al., 2024; Kittredge, Lee, & Jiang, 2025)

ID	항목	측정	가중	판정	근거·출처
B5	교수설계·인지부하	HYBRID	권장	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
B6	교수 통합·교수자 지원	DOC	권장	○	Duolingo Max 는 Super(Super Duolingo) 혜택에 AI 기능을 더한 등급인데 '듀오링고 교실'에 가입한 상태에서는 Super 를 쓸 수 없고 교실을 탈퇴해야 구독을 살 수 있다. Duolingo for Schools 의 교사 설정은 친구 요청·성인 콘텐츠 표시를 켜고 끄는 것으로 안내되며 신규 계정·교실을 만들 수 없고 2027 년 7 월 31 일 종료될 예정이다. Duolingo Max AI 기능을 수업에 통합하는 교사 자료나 교육기관용 Duolingo Max 안내는 문서로 확인되지 않아 미충족으로 판정했다. (출처: Duolingo, Retrieved 2026-a; Duolingo, Retrieved 2026-b; Duolingo, 2026g; Duolingo, 2026f)
B7	상호작용 사용성·접근성(교육 맥락)	HYBRID	보조	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음). 참고: 공식 도움말은 Duolingo Max 를 웹에서는 이용할 수 없고 모바일 앱에서만 이용 가능하다고 안내하며 Explain My Answer·Roleplay 제공 과정 목록에는 한국어 사용자용 영어 과정이 없다. (출처: Duolingo, Retrieved 2026-a)
B8	표준 준수·상호운용	DOC	보조	○	1EdTech 인증 등록부에서 Duolingo 제품의 인증 기록은 확인되지 않았고, LTI 등 교육 데이터·콘텐츠 상호운용 표준으로 Duolingo Max 기능을 연동하는 방법도 문서로 확인되지 않아 미충족으로 판정했다. (출처: 1EdTech Consortium, Retrieved 2026)
B9	효과성 증거 등급	DOC	필수	○	제공사 독립·고등교육 대상·제품 단위(Duolingo Max AI 기능) 조건을 갖춘 연구가 없어 독립 효과 근거 없음으로 판정했다. Duolingo Max 를 다룬 제공사 독립 고등교육 연구 2 편은 Max 를 다른 앱·AI 도구와 묶어 처치해 제품 단위 효과를 가를 수 없고 단일집단 설계·측정 도구의 낮은 신뢰도 등 연구 질의 한계가 있어 반영하지 않았다. 참고: Duolingo 소속 연구팀의 Video Call 무작위 통제 실험(일본어 화자 567 명, 자체 보고, 동료심사 전)은 30 일 이용 뒤 사후 말하기 점수가 회귀 분석에서 통제 조건보다 2 점 높은 작은 정적 효과를 보고했다. 같은 소속 연구팀의 학술지 논문(385 명)은 통제집단 없이 자기효능감 상승을 보고했고 제공사와 독립된 중국 대학 1 학년 80 명 준실험은 Duolingo Max 기능이 아닌 일반 Duolingo 앱을 다뤘다. (출처: Kittredge, Lee, & Jiang, 2025; Kittredge, Hopman, et al., 2025; Ouyang et al., 2024)

ID	항목	측정	가중	판정	근거·출처
B10	효과크기 정량	DOC	권장	○	제공사 독립·고등교육 대상·제품 단위(Duolingo Max AI 기능) 조건을 갖춘 연구가 없어 효과크기를 기록하지 못해 미충족으로 판정했다. Duolingo Max 를 다룬 제공사 독립 고등교육 연구 2 편은 Max 를 다른 앱·AI 도구와 묶어 처치해 제품 단위 효과를 가를 수 없고 단일집단 설계·측정 도구의 낮은 신뢰도 등 연구 질의 한계가 있어 반영하지 않았다. 참고: 자체 보고(동료심사 전) Video Call 실험의 회귀 분석 차이는 말하기 점수 2 점으로 보고서 스스로 작은 정적 효과로 규정했다. 일반 Duolingo 앱을 다룬 독립 준실험의 학습 참여 총점 효과크기는 Cohen's d = 2.03 이었으나 Duolingo Max 기능 대상이 아니다. (출처: Kittredge, Lee, & Jiang, 2025; Ouyang et al., 2024)
C. AI 신뢰성·윤리·거버넌스 (가중치 30%)					
C1	타당성·신뢰성과 환각 관리	HYBRID	필수*	미측정	(게이트 항목) 실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음). 참고: Duolingo 는 AI 생성 내용의 사실 정확성·어조를 계속 검토하고 학습자가 부정확한 응답을 신고할 수 있다고 밝힌다. 연차보고서는 생성형 AI 가 부정확한 내용과 환각을 낼 수 있고 정확성 조치가 늘 성공하지는 않아 이용자 신고에 의존할 수 있다고 공시했다. (출처: Duolingo, 2023; Duolingo, 2026b)
C2	투명성·설명가능성	DOC	필수	◐	Duolingo 블로그와 도움말은 Explain My Answer·Roleplay·Video Call 을 생성형 AI 기반 기능으로 소개하고 Roleplay 상대가 실제 사람이 아니라고 안내하며 개인정보처리방침은 AI 기능 입력이 OpenAI·Google 같은 공급업체와 공유될 수 있다고 밝힌다. Explain My Answer 는 학습자의 오류와 해당 규칙이 어려운 이유를 고려해 설명하지만 Video Call·Roleplay 응답의 근거 제시와 기능 화면 안의 AI 고지, 기반 모델의 현행 명칭은 문서로 확인되지 않아 부분 충족으로 판정했다. (출처: Duolingo, 2023; Duolingo, Retrieved 2026-a; Duolingo, 2026c; Mas Castillo, 2026)
C3	공정성·편향 관리	HYBRID	필수	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).

ID	항목	측정	가중	판정	근거·출처
C4	프라이버시·데이터 거버넌스	DOC	필수	◐	개인정보처리방침은 AI 기능 입력을 OpenAI·Google 같은 공급업체와 공유할 수 있되 업체의 자체 목적 사용을 계약으로 막는다고 하고 아동 사용자 별도 취급, 미국 내 처리, Data Vault(개인정보 열람·삭제 요청 도구)를 통한 접근·삭제 요청을 고지한다. 그러나 AI 대화의 녹음·전사본을 자체 AI 모델 학습에 쓸 수 있고 개인정보는 일반적으로 계정 삭제 때까지 보관되며 학습 활동 데이터 권리 양도, 학교와의 학생 데이터 프라이버시 협약(SDPA) 미체결, 한국 이용자 전용 조항 부재가 확인되어 부분 충족으로 판정했다. (출처: Duolingo, 2026c; Duolingo, 2025; Duolingo, 2026g)
C5	인간 통제·교수자 역할	DOC	필수*	◐	(게이트 항목) Roleplay 시나리오와 첫 메시지는 전문가가 써서 모델에 대화 방향을 지시하고 Duolingo 가 AI 생성 내용을 계속 검토하며 학습자는 부정확한 응답을 신고할 수 있으나 연차보고서는 정확성 조치가 늘 성공하지는 않아 이용자 신고에 의존할 수 있다고 공시했다. '듀오링고 교실'에 가입하면 Super 를 쓸 수 없고 Duolingo for Schools 의 교사 설정은 친구 요청·성인 콘텐츠 표시로 안내되며 신규 교실을 만들 수 없다. Duolingo Max AI 기능에 교수자가 개입하는 경로가 확인되지 않아 부분 충족으로 판정했다. (출처: Duolingo, 2023; Duolingo, 2026b; Duolingo, Retrieved 2026-b; Duolingo, 2026g; Duolingo, 2026f)
C6	안전·견고성·보안	HYBRID	권장	미측정	실사용 측정이 필요한 항목이어서 공개 문서만으로는 판정하지 않았다(실사용 진단 결과는 이번 호에 싣지 않음).
C7	책무성·거버넌스·정책 정합	DOC	권장	◐	개인정보 문의는 데이터 보호 책임자에게, 접근·삭제 요청은 Data Vault 로 하도록 안내하고 학습자가 부정확한 AI 응답을 신고할 수 있으며 연차보고서는 생성형 AI 의 부정확성·환각·편향 위험을 공시했다. 서비스 약관은 정확성과 사용 성과를 보증하지 않고 출신 국가와 관계없이 미국 펜실베이니아주 법을 준거법으로 하며 금전 분쟁은 영어 개별 중재(최초 이용 후 30 일 안 탈퇴 가능)로 해결하고 총 책임을 직전 12 개월 결제액으로 제한한다. 한국 이용자 전용 조항·국내대리인 등 국내 정책 대응도 확인되지 않아 부분 충족으로 판정했다. (출처: Duolingo, 2026c; Duolingo, 2023; Duolingo, 2026b; Duolingo, 2025)

ID	항목	측정	가중	판정	근거·출처
C8	AI 리터러시·연령 적합성	DOC	보조	◐	개인정보처리방침은 미국 기준 13 세 미만이거나 각국 디지털 동의 연령 미만인 사용자를 아동 사용자로 따로 취급해 부모 통제와 최소 수집을 적용한다고 밝힌다. 블로그는 Roleplay 상대가 실제 사람이 아니라고 안내하고 부정확한 AI 응답을 신고하는 방법을 알리지만 Duolingo Max AI 기능의 이용 연령 기준과 학습자용 AI 리터러시 자료는 문서로 확인되지 않았다. 연령별 처리 원칙과 AI 사용 안내만 확인되어 부분 충족으로 판정했다. (출처: Duolingo, 2026c; Duolingo, 2023)
축 점수 A 미산출 · B 19.2 · C 50.0 → 종합 32.4 | 고부담 게이트 해당 없음 | 도입 검토 의견 판정 유보					

자료 기준일까지 공개된 문서로 26 개 항목 중 11 개(필수 13 개 중 5 개)를 판정했다. 실사용 측정이 필요한 14 개 항목(게이트 B3·C1 포함)과 만족도 근거를 찾지 못한 A8 은 미측정으로 점수에서 뺐다. 축 A 는 판정한 항목이 없어 미산출이고 종합 32.4 점은 축 B·C 의 가중치(40·30)로만 계산했다. 필수 항목 13 개 가운데 판정한 항목이 5 개로 경계값 적용 최소 범위(10 개)에 미치지 못해 도입 검토 의견은 '판정 유보'다. 게이트 항목 C5 가 ◐이고 필수 항목 B9 가 ○여서 이 판정들이 바뀌지 않는 한 미측정 항목을 측정해도 적극 검토에는 이를 수 없다. 공통 항목 기준 점수(32.4 점)는 판정한 11 개 항목이 모두 공통 항목이어서 종합 점수와 같다. 축 A 가 CS50 Duck 에만 있으므로 두 제품의 종합 점수끼리는 직접 비교하지 않는다. 축 점수는 소수 첫째 자리까지 반올림하고 종합 점수는 반올림한 축 점수로 계산했다.

주: ●(충족)·◐(부분 충족)·○(미충족)·미측정으로 판정하고 미측정은 점수에서 제외했다. 측정 열의 AUTO·HYBRID 는 실측해야 판정하는 항목이어서 두 제품 모두 미측정이고 DOC 는 두 제품 모두 자료 기준일까지 공개된 공식 문서와 연구로만 판정했으며 B9·B10 에는 제공사와 독립적으로 수행된 고등교육 대상 연구만 반영했다. 가중 열의 별표(*)는 게이트 항목(B3·C1·C5)이며 고부담 게이트는 이번 호에 적용하지 않았다. 필수 항목 13 개 가운데 측정이 10 개 미만이면 판정 유보로 적으며 이는 결함 판정이 아니다. 두 제품의 비교에는 공통 측정 항목 11 개의 점수만 쓴다.

부록 3. 용어 해설

용어	설명
AI 튜터(대화형 튜터)	특정 교과·강좌·기능 학습을 위해 교수법을 설계한 학습자 대면 대화형 AI. 정답을 곧바로 주기보다 질문·단계적 힌트·피드백으로 학습을 이끄는 것을 목표로 하며 이 보고서의 분석 단위다
대규모 언어 모델(LLM)	방대한 텍스트로 학습해 질문에 답하거나 글을 생성하는 AI 모델. 이 보고서가 다루는 AI 튜터 대부분이 LLM 을 기반 모델로 쓴다(Ⅱ 장 1 절)
지능형 튜터링 시스템(ITS)	학습자의 응답을 분석해 풀이 단계별 힌트와 피드백을 주는 컴퓨터 튜터. LLM 이전 튜터 연구의 비교 기준이다(Ⅱ 장 3 절 가)
검색 증강 생성(RAG)	강의 자막·교재 같은 외부 자료에서 관련 내용을 찾아 LLM 의 답에 넣는 방식. 근거 제시에 쓰이며 자료 이용 범위의 저작권 검토가 따른다(Ⅳ 장 1 절 가)
시스템 프롬프트·가드레일	시스템 프롬프트는 서비스가 LLM 에 미리 넣어 두는 역할·규칙 지시문이고 가드레일은 정답을 곧바로 주거나 강좌와 무관한 요청에 답하는 것처럼 원치 않는 응답을 막으려고 덧댄 지시·필터·점검 장치다(Ⅱ 장 1 절 나)
학습관리시스템(LMS)·LTI	LMS 는 강좌 자료·과제·성적을 관리하는 대학의 온라인 강의 플랫폼(Moodle·Canvas 등)이고 LTI(학습 도구 상호운용성)는 외부 학습 도구를 LMS 안에서 쓰도록 연동하는 표준이다(Ⅰ 장 1 절 나, Ⅱ 장 1 절 마)
스캐폴딩(비계)	학습자가 혼자 하기 어려운 과제를 단계적 힌트·질문으로 돕고 익숙해지면 지원을 줄여 가는 교수 설계(Ⅱ 장 1 절 라, Ⅱ 장 3 절 나)

용어	설명
소크라테스식 질문	답을 알려 주기보다 질문을 던져 학습자가 스스로 추론하게 하는 대화 방식
오개념 진단	학습자의 답이나 풀이에서 잘못된 개념 이해를 찾아내는 일. 잘못된 추론으로 정답에 이르는 '정답 함정'은 놓치기 쉽다(Ⅱ장 1절 다)
정답 노출과 교수법적 탈옥	튜터가 지시와 달리 완성된 답이나 코드를 주는 현상과 학습자가 역할극·우회 요청으로 답을 끌어내려는 시도(Ⅱ장 4절 나)
가상 조교(VTA)	강좌 자료를 근거로 학생 질문에 답하는 AI 조교. 자격기준을 대통령령으로 정하는 고등교육법상 조교와는 다른 명칭이다(Ⅳ장 1절 다)
신기성 효과	새 도구에 대한 초기 흥미로 단기 효과가 크게 나타났다가 시간이 지나며 줄어드는 현상(Ⅱ장 4절 가)
무작위 통제 실험(RCT)과 교차설계	RCT는 참여자를 처치 조건과 비교 조건에 무작위로 배정해 효과를 견주는 실험이고 교차설계는 같은 참여자가 시기를 달리해 두 조건을 모두 거치게 하는 방식이다(Ⅱ장 3절 나)
효과크기(d·g·ES·SD)	두 조건의 평균 차이를 표준편차 단위로 나타낸 값. 원문 표기에 따라 d·g(표준화 평균 차이), ES(메타분석이 보고한 효과크기), SD(표준편차 단위로 적은 차이)로 쓰며 비교 조건이 다르면 같은 크기라도 직접 견주기 어렵다(Ⅱ장 3절)
K-12	초·중등 교육(유치원부터 12학년까지). 이 보고서에서는 대학 강좌가 아닌 학교급의 연구임을 알리는 근거 표지로 쓴다(Ⅰ장 1절 가)
근거 표지	근거의 성격을 문장 안 괄호로 밝히는 표시. 동료심사 전(학술지·학술대회의 동료심사를 거치지 않은 워킹페이퍼·프리프린트·기술 보고서), 개발팀 수행(평가 대상 튜터를 연구자가 직접 개발·운영한 연구), 자체 보고(제공사 수행·후원), 벤치마크·시뮬레이션(실제 학생 없이 문항 세트나 가상 학생으로 한 평가)이 있다. 근거 수준 정의는 [표 Ⅱ-3] 주를 따른다(Ⅰ장 4절)
게이트 항목과 판정 유보	게이트 항목(B3·C1·C5)은 도입 검토 의견의 상한을 정하는 평가척도 항목으로 적극 검토에는 모두 ●, 조건부 검토에는 모두 ◐ 이상이 필요하다. 판정 유보는 필수 항목 13개 가운데 측정한 항목이 10개 미만이어서 경계값을 적용하지 않았다는 뜻이며 결함 판정이 아니다(Ⅰ장 4절, Ⅲ장 2절 다)

부록 4. 기관 도입 점검표

본문의 점검 항목을 단계별로 모았으며 담당은 예시다.

단계	점검 항목	담당(예시)	관련 절
계약 전	강좌 단위 수요와 학생이 이미 쓰는 범용 도구 대비 차별점(강좌 자료 근거·정답 보류)을 확인했는가	교무·교수학습개발센터	Ⅱ장 2절 가·마, Ⅳ장 1절 라
계약 전	제공사와 독립된 효과 연구가 있고 과목·인지 수준·기간·학습자 특성이 학내 강좌와 비슷한가	교무·교수학습개발센터	Ⅱ장 3절, Ⅱ장 4절 가
계약 전	정답 보류를 응답 사후 점검·순차 안내 같은 구조로 강제하는지 공급사·개발팀에 확인했는가	교수학습개발센터·AI에듀테크센터	Ⅱ장 1절 나·라
계약 전	대화 기록의 보관 기간, 모델 학습 등 2차 이용, 위탁·국외 처리 조건을 계약 문서로 확인하고 가명처리 기준을 정했는가	개인정보 보호·정보화	Ⅲ장 1절, Ⅳ장 1절 나
계약 전	강좌 자료 연결의 수업 목적 이용 범위·이용 허락·프로그램 특례와 수강생 외 접근 제한을 확인했는가	저작권 담당·교수자	Ⅳ장 1절 가

단계	점검 항목	담당(예시)	관련 절
계약 전	LMS 연동 범위, 평가 기간 지원 체계, 등급·가격·환불 조건과 서비스 종료 같은 지속성 조건을 확인했는가	정보화·교무	Ⅱ장 2 절 라, Ⅳ 장 1 절 라
도입 설정	정답 직접 요구·오개념·허위 전제·근거 요청 과업을 한국어 자료로 실행해 정답 보류와 내용 정확성을 실측했는가	교수학습개발센터·AI 에듀테크센터	Ⅱ장 1 절 나, Ⅱ 장 4 절 바, Ⅲ 장 2 절
도입 설정	음성 기능의 발음 피드백 정확성과 감정 추론 기능 여부(규제 경계)를 확인했는가	교수학습개발센터·정보화	Ⅱ장 1 절 마, Ⅱ 장 4 절 사
도입 설정	화면의 AI 사용 고지와 이용 안내(오류 가능성·출처 확인·신고 경로), 이용 연령 기준을 갖췄는가	교무·정보화	Ⅱ장 4 절 다·사, Ⅳ 장 1 절 가
도입 설정	강좌 정책으로 허용 AI 의 범위와 AI 없이 치르는 평가를 정했는가	교수자	Ⅱ장 4 절 나, Ⅲ 장 1 절
도입 설정	전면 도입 전에 대상 강좌·기간·확대 판단 기준을 정한 소규모 시범 운영을 두었는가	교무·교수학습개발센터	Ⅱ장 4 절 라, Ⅳ 장 2 절 나
수업 운영	성적 관련 질문은 사람이 답하고 AI 응답은 조교가 검수하도록 역할을 나누고 그 업무를 공식화했는가	교수자·조교	Ⅳ장 1 절 다, Ⅳ 장 2 절 나
수업 운영	튜터를 수업 활동과 결합하고 사용 목적을 안내했는가	교수자	Ⅱ장 4 절 라, Ⅳ 장 2 절 나
수업 운영	의인화·정서적 기능의 수준을 점검하고 자기점검 전략을 병행하는가	교수자·학생지원	Ⅱ장 4 절 마
점검·평가	기반 모델이 바뀔 때 같은 과업으로 다시 점검하는가	AI 에듀테크센터	Ⅲ장 3 절, Ⅳ 장 2 절 나
점검·평가	AI 없이 치른 평가·지연 측정을 둔 학내 효과 측정과 개인 식별 없는 사용률 점검을 설계하고 운영 관계가 있는 서비스는 관계없는 연구자가 분석하는가	교무·AI 에듀테크센터	Ⅱ장 4 절 가, Ⅳ 장 2 절 나
점검·평가	학습자 집단별 사용률·성과 차이를 점검하는가	교무·AI 에듀테크센터	Ⅱ장 4 절 사

PUBLICATION INFORMATION

발간 정보

항목	내용
보고서명	월간 에듀테크 솔루션 분석 보고서 Vol.03 AI 튜터링 플랫폼
대상 월	2026 년 8 월분 (전 8 호 중 제 3 호)
발간일	2026. 8. 31.
자료 기준일	2026. 8. 31. — 제품의 기능·가격·정책 정보는 이 날짜 현재 기준임
발행처	한림대학교 AI 에듀테크 연구과제
과제	글로컬대학 30 한림대학교 AI 에듀테크센터 연구과제「고등교육 AI 에듀테크의 통합적 발전 방안 연구 — 전공 운영 모델·연구 방향성 정립·기초교양 AI 리터러시 교육」[연구 2] 월간 에듀테크 솔루션 분석
연구책임자	김슬기 (한림대학교 일송자유교양대학)
집필진	김슬기(연구책임자)·고영웅·김진환·김병정·김성우·김용수·서동기·육청민·이정근·정혜선·진서연·한민정·한수미·홍수현·황현석 (이상 한림대학교)
문의	skkim@hallym.ac.kr

인용 표기

김슬기·고영웅·김진환·김병정·김성우·김용수·서동기·육청민·이정근·정혜선·진서연·한민정·한수미·홍수현·황현석(2026). 월간 에듀테크 솔루션 분석 보고서 Vol.03: AI 튜터링 플랫폼. 한림대학교 AI 에듀테크 연구과제.

이용 안내

○ 이 보고서의 저작권은 한림대학교에 있습니다. 출처를 밝히면 비영리 목적으로 인용·활용할 수 있습니다.

○ 이 보고서에서 다룬 제품·서비스는 분석 목적으로 선정한 것이며 특정 제품의 홍보나 보증을 뜻하지 않습니다.

○ 제품의 기능·가격·정책은 자료 기준일(2026. 8. 31.) 이후 바뀔 수 있습니다.

○ 이 보고서의 내용은 연구진의 연구 결과이며 한림대학교 또는 AI 에듀테크센터의 공식 견해와 다를 수 있습니다.

○ 이 보고서는 HAI-GPT 를 운영하는 한림대학교 AI 에듀테크센터의 연구과제 산출물로, 이 센터가 운영·실증하는 HAI-MathTutor(AI 선형대수 튜터)·HAI-Tutor·K-University 'AI 동반자'와 HAI-GPT 스터디 튜터는 평가하지 않고 사례로만 다루었습니다. 연구진에는 해당 부서·센터 구성원이 포함됩니다. 이 보고서가 게시되는 KELI 플랫폼의 솔루션 마켓에는 센터가 운영하는 HAI-MathTutor(솔루션 공급 기업 슈퍼러닝)가 함께 소개되어 있습니다. AI 선형대수 튜터는 한림대학교가 에듀테크 기업 슈퍼러닝과 함께 개발했다고 발표한 서비스입니다. 채점하지 않은 서비스에 결함 근거가 드러나지 않은 것을 상대적 우위로 해석해서는 안 됩니다.

사사

본 과제(결과물)는 2026 년도 교육부 및 강원특별자치도의 재원으로 강원 RISE 센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE) 글로컬대학 30 의 결과입니다.(2026-RISE-10-009)

This research was supported by the Regional Innovation System & Education(RISE) Glocal University 30 Project program through the Gangwon RISE Center, funded by the Ministry of Education(MOE) and the Gangwon State(G.S.), Republic of Korea.(2026-RISE-10-009)

MONTHLY EDTECH SOLUTION ANALYSIS REPORT

월간 에듀테크 솔루션 분석 보고서 시리즈

호	대상 월	주제
Vol.01	2026 년 6 월분	LLM 기반 학습 도구
Vol.02	2026 년 7 월분	적응형 학습 시스템
Vol.03	2026 년 8 월분	**AI 튜터링 플랫폼**
Vol.04	2026 년 9 월분	학습분석 도구
Vol.05	2026 년 10 월분	AI 평가·피드백 도구
Vol.06	2026 년 11 월분	인간-AI 공동조절 학습 지원 도구
Vol.07	2026 년 12 월분	교수자용 생성형 AI 도구
Vol.08	2027 년 1 월분	AI 거버넌스·학업 진실성 솔루션
종합	2027 년 1 월	종합 동향 리포트

이번 호 소개

「월간 에듀테크 솔루션 분석 보고서」 Vol.03(2026 년 8 월분)은 교수법을 설계한 대화형 AI 튜터를 기술·시장·교육적 적용·한계의 네 측면에서 분석했다. 정답을 곧바로 주지 않는 지시는 프롬프트만으로 지켜지기 어려웠고 이를 보완하는 구조 설계는 연구 단계가 많았다. 대학 강좌 실험의 효과는 비교 조건·측정·설계에 따라 크게 엇갈렸다. CS50 Duck 과 Duolingo Max 는 공개 문서로 평가해 필수 항목을 5 개만 판정할 수 있어 판정 유보였다. 대학은 교수자·조교의 역할을 나누고 저작권·데이터·고지 조건을 도입 전에 점검할 수 있다.

키워드 AI 튜터 · 고등교육 · 교수법 설계 · 정답 노출 · 가상 조교

한림대학교 AI 에듀테크 연구과제

글로컬대학 30 한림대학교 AI 에듀테크센터 연구과제「고등교육 AI 에듀테크의 통합적 발전 방안 연구」[연구 2] 산출물