

LearnLM: 학습을 위한 Gemini 개선

LearnLM 팀, Google

현재의 생성형 AI 시스템은 기본적으로 정보를 제시하도록 조정되어 있으며, 인간 튜터처럼 학습을 위해 사용자와 상호작용하지 않습니다. 이러한 시스템의 광범위한 잠재적 교육 활용 사례를 해결하기 위해, 우리는 교육적 행동을 주입하는 과제를 '교육적 지시/파르기'로 재구성합니다. 여기서 훈련 및 평가 예시는 후속 모델 실행에 존재하거나 요구되는 특정 교육학적 속성을 설명하는 시스템 수준의 지시를 포함합니다. 이러한 접근법은 모델을 특정 교육학적 정의에 얹매이지 않게 하며, 대신 교사나 개발자가 원하는 모델 행동을 명시할 수 있게 합니다. 또한 훈련 후 혼합에 교육학적 데이터를 추가할 수 있도록 함으로써, 급속히 확장되는 기능 세트와 병행하여 Gemini 모델의 학습 능력을 향상시키는 길을 열어줍니다. 두 가지 모두 초기 기술 보고서^[1] 대비 중요한 변화입니다. 교육적 지시문 추종 훈련을 통해 생성된 LearnLM 모델(Google AI Studio에서 이용 가능)이 다양한 학습 시나리오 전반에 걸쳐 전문가들로부터 현저한 선호도를 얻었음을 보여줍니다. 평균 선호도 강도는 GPT-4o 대비 +31%, Claude 3.5 Sonnet 대비 +11%, LearnLM의 기반이 된 Gemini 1.5 Pro 모델 대비 +13%를 기록했습니다.

1. 서론

2024년 5월 발표된 당사의 [초기 기술 보고서\[1\]](#)는 교육 기술의 역사와 현재 현황을 조사하고, 생성형 인공지능(생성 AI)이 교육에 미칠 잠재적 영향을 논의하며, 평가 개발을 위한 당사의 협력적 접근 방식을 제시했습니다.

보고서 발간 후, 학교, 교육 기술(에드테크) 기업, 비영리 단체, 정부 기관 등 국제 교육 분야의 다양한 주체로부터 모델 적용 또는 협업 의사를 포함한 의견이 접수되었습니다. 이러한 제출 자료 검토, 20회 이상의 후속 인터뷰, 생성형 AI 기반 학습 기능을 개발 중인 구글 제품팀의 의견을 종합하여 주요 결과를 다음과 같이 요약합니다:

1. 교육학^[1]또는 더 정확히 말해 AI 튜터의 이상적인 행동 양상은 수용해야 할 학년 수준, 과목, 언어, 문화, 제품 설계, 철학 등이 너무 광범위해 정의하기가 매우 어렵다. 공통점이 많긴 하지만, 서로 다른 맥락에서 적절한 행동은 다를 수 있으며 심지어 상충할 수도 있으므로 개발자나 교사가 구체적으로 명시하는 것이 가장 바람직하다.
2. 인공지능 학습 시스템을 개발할 때, 기본 모델에서 가장 흔히 언급되고 즉시 활용 가능한 기능은 시스템 지시를 따라 상호작용형 튜터 주도형 연습 문제를 생성하는 능력이다. 이러한 지시를 명시하는 교사나 개발자는 학생이 이를 회피하려 시도하더라도(예: "정답을 알려주지 마라" 또는 "주제를 벗어나지 마라") 인공지능 튜터가 지정된 지시를 정확히 따를 것이라고 확신하고 싶어 한다.
3. 각 애플리케이션에 대한 사후 미세 조정은 단기적으로는 효과적일 수 있으나, 비용, 유지 관리, 그리고 빠르게 발전하는 기본 모델로 인해 실용적이지 않습니다. 따라서 단점이 있음에도 불구하고 프롬프팅은 교육 제품 개발자가 행동을 지정하는 최선의 방법으로 남을 가능성성이 높습니다.

본 논문은 다음과 같은 관점에서 모델링 및 평가 방법론을 업데이트한 과정을 설명한다

¹ 우리는 교육학(pedagogy)이라는 용어를 가능한 한 광범위한 의미로 사용하며, 어린이 교육에 국한되지 않고 인간의 교수법 및 관련 학습 기법을 포함합니다.

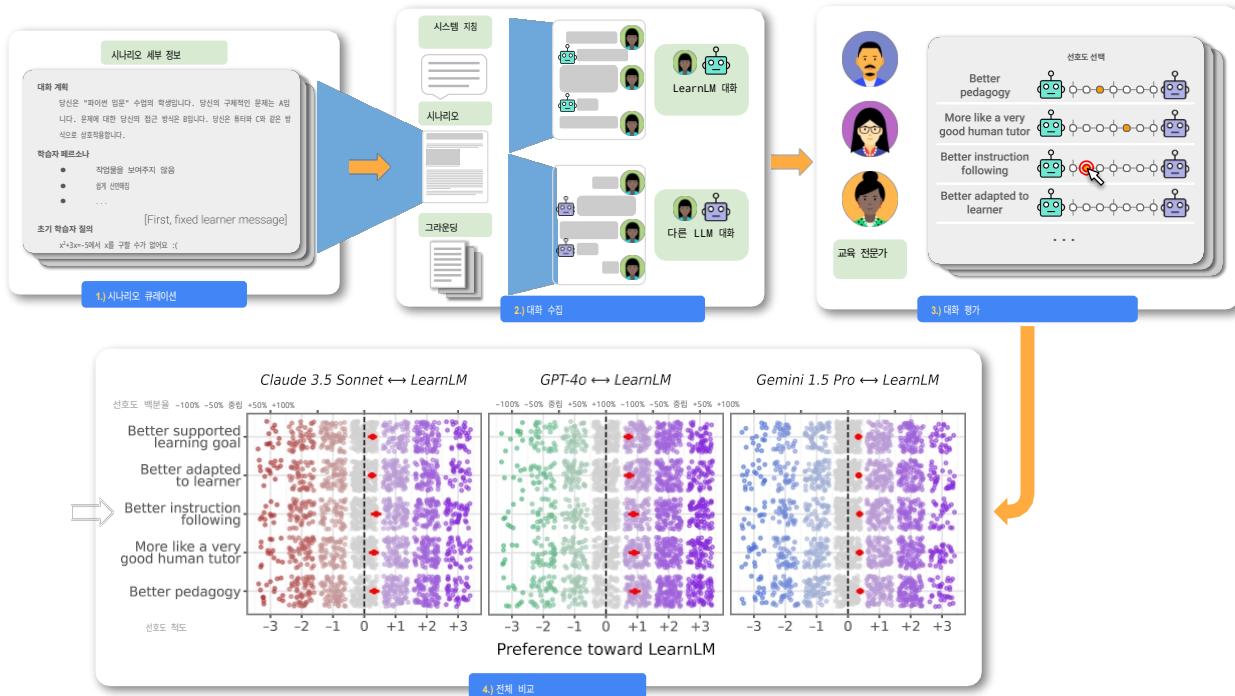


그림 1 | LearnLM과 다른 시스템을 비교하기 위한 3단계 전문가 평가 파이프라인 개요 및 결과. (1) 전문가 참가자들이 AI 튜터 쌍과 상호작용하는 특정 학습자 역할을 수행할 수 있는 학습 시나리오를 개발했습니다. (2) 각 시나리오에 특화된 기초 자료(예: 에세이, 숙제 문제, 도표 등)와 시스템 지침을 각 모델에 컨텍스트로 전달합니다. 생성된 대화 쌍을 교육학 전문가에게 전달하여 (3) 각 모델의 개별 성능과 비교 성과를 검토하게 합니다. 이러한 비교 평가(7점 응답 척도)를 집계하여 (4) LearnLM이 GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro 대비 전반적인 선호도를 평가합니다. 자세한 결과는 [4절](#)을 참조하십시오.

이러한 관찰 사항들입니다. 구체적으로, 우리는 본 연구를 '추종하는 교육적 지침'으로 설정했습니다. 즉, 훈련 및 평가 예시에 시스템 수준의 지침을 맥락화하여 원하는 교육적 행동을 설명합니다. 이 접근법은 시스템의 행동 방식을 좁게 규정하지 않으며, 인격이나 스타일의 충돌 없이 Gemini의 나머지 훈련 혼합물에 교육적 데이터를 효과적으로 추가할 수 있게 합니다. 또한 모델이 보다 미묘한 교육적 지침과 선호도를 따를 수 있도록 인간 피드백 기반 강화 학습(RLHF) [2]를 훈련 절차에 포함시켰습니다.

업데이트된 방법론을 사용하여 Gemini 1.5 Pro²[3] 기반의 LearnLM 새 버전을 훈련했습니다. 2024년 10월 1일 기준 각 기업의 최상위 모델을 대표하는 동시대 플래그십 모델들과의 평가에서, 교육자와 교육학 전문가들은 이 버전의 LearnLM을 선호했으며, 평균 선호도는 GPT-4o 대비 +31%, Claude 3.5 Sonnet 대비 +11%, 기존 Gemini 1.5 Pro 대비 +13% 높았습니다(그림 1 참조). LearnLM은 [Google AI Studio](#)에서 실험용 모델로 제공되며, 사용 사례 예시와 권장 프롬프트 문서가 함께 제공됩니다. 향후 연구 및 우선순위 설정에 도움이 될 LearnLM에 대한 [피드백](#)을 환영합니다. 교육 및 학습을 위한 LearnLM 개선과 동시에, 이러한 발전 사항을 Gemini 모델에도 적용하기 위해 노력하고 있습니다. 따라서 Gemini를 사용하는 모든 개발자는 LearnLM 연구를 통해 이루어진 개선 사항의 혜택을 누릴 수 있습니다.

[제2절에서는](#) 교육적 지도를 위해 LearnLM을 훈련한 방법을 설명하고, [제3절에서는](#) 이에 따라 시나리오 기반 평가 설계를 업데이트한 과정을 설명합니다. [제4절에서는](#) LearnLM을 다른 주요 모델과 비교한 결과에 대한 상세한 분석을 제시합니다. 마지막으로 [제5절에서는](#) 특히 지속적인 평가와 관련된 향후 연구 방향을 개괄합니다. 광범위한 범위 외에도

² 특히 gemini-1.5-pro-002 ([릴리스 노트](#)).

훈련 및 평가에 사용한 핵심 학문 분야 외에도, [부록 C에는](#) 의학 교육 분야에 대한 타당성 연구를 포함시켰습니다.

2. 모델링

원본 기술 보고서[1]에서 우리는 다양한 합성 및 인간 작성 데이터셋을 활용한 지도형 미세 조정(SFT)을 통해 기본 모델의 행동을 조정했습니다. 이후 훈련 전략에 상당한 변경을 가했습니다: 첫째, 교육적 지시 추종에 중점을 두어 SFT 데이터를 업데이트했습니다. 둘째, 인간 피드백 기반 강화 학습(RLHF)[2]을 추가로 활용하기로 결정했습니다. 이를 위해 인간 선호도 데이터를 수집하여 보상 모델(RMs)을 훈련시키고 RL 단계를 위한 프롬프트를 생성했습니다. 셋째, Gemini의 표준 사후 훈련 후 자체 사후 훈련을 수행하기보다는 Gemini와 공동 훈련을 진행했습니다. 즉, 우리의 데이터를 Gemini의 SFT, RM, RL 단계에 직접 혼합했습니다. LearnLM은 이러한 실험적 혼합의 결과물이며, 우리는 또한 우리의 데이터와 평가를 주요 Gemini 모델에 통합해 왔습니다. LearnLM의 개선 사항 중 일부는 최근 출시된 [Gemini 2.0 모델\[4\]](#)에 포함되어 있습니다.

2.1. 교육적 지시 따르기

지시 따르기(IF)는 모델이 프롬프트를 따르는 능력을 의미하며, 일반적으로 인간의 의도와 더 잘 부합하기 위함이다[5]. Gemini[3]는 대화 중 사용자가 삽입하는 사용자 지시와 개발자가 사용자 상호작용 전에 미리 지정하는 [시스템 지시](#)를 구분하며, 시스템 지시는 이후 사용자가 제공하는 모든 지시보다 우선한다. 시스템 지침은 복잡성이 크게 달라질 수 있습니다. "당신은 지식이 풍부한 글쓰기 코치입니다"와 같이 최소한으로 명시된 단일 문장에서부터, "사용자가 3개의 질문에 정답을 맞췄다면 다음 주제로 이동하라"와 같은 특정 조건부 기대치, 또는 Mollick과 Mollick[6]의 교육 프롬프트나 최근 제안된 복합적 IF 벤치마크[7]에서 볼 수 있는 복잡한 작업과 행동을 설명하는 상세한 여러 단락의 지침에 이르기까지 다양합니다.

지침은 크게 두 가지 범주로 나뉩니다: 하드 제약 조건은 주로 길이, 서식 또는 콘텐츠 요구 사항에 사용되며(예: "100단어 미만으로 텍스트 요약하기" 또는 "단어 X 사용 금지"), 소프트 제약 조건은 스타일, 인격 또는 어조를 제어하는 데 사용되는 더 미묘한 제약 조건이나 가이드라인입니다(예: "전문적인 어조 사용하기" 또는 "비원어민이 이해하기 쉬운 언어 사용하기"). 오픈소스 인터랙티브 픽션(IF) 벤치마크 중 IFEval[8]은 하드 제약의 하위 집합인 프로그램적으로 검증 가능한 IF에 초점을 맞추고 있으며, Qin 등[9]과 같은 최근 벤치마크는 더 미묘한 언어적·스타일적 지침을 포함하도록 범위를 확장하고 있습니다. 교육적 사용 사례에서는 두 범주의 지침 모두 중요합니다. 예를 들어 "정답을 밝히지 마라"는 하드 제약인 반면, "동기를 부여하는 어조를 사용하라"는 소프트 제약입니다.

IF(지시 따르기) 기능의 개선은 이미 다양한 학습 사용 사례에서 더 나은 모델 응답을 이끌어냈습니다. 본 연구에서는 이러한 진전을 바탕으로, 교육적 시스템 지시에 대한 지시 따르기 성능 향상에 집중합니다. 교육적 시스템 지시는 일반적으로 더 복잡하고 미묘하며 검증하기 어려운 특성을 지닙니다. 이러한 특성들로 인해 모델이 이를 따르기 더 어렵습니다.

2.2. 훈련 후 데이터 수집 전략

주요 모델링 전략은 AI 튜터 개발자들이 흔히 사용하는 교육적 시스템 지시를 모델이 더 잘 따르도록 하는 데이터를 수집하는 것입니다. 이에 따라 SFT 데이터를 업데이트하여 각 대화마다 해당 대화에서 나타나는 교육적 행동을 구체적으로 설명하는 서로 다른 시스템 지시로 시작하도록 했습니다. 더 일반적이거나 모호한 지시는 역효과를 납니다. 모델이 목표 모델의 회전을 예측하는 데 유용하지 않은 지시를 무시하도록 학습하기 때문입니다.

목표 모델의 응답을 예측하는 데 유용하지 않은 지시를 무시하도록 학습하기 때문입니다.

인간 선호도 데이터를 수집하기 위해, 우리는 마찬가지로 각 대화에 서로 다른 교육적 초점을 둔 시스템 지침을 시드(seed)로 투입하고, 평가자들에게 모델 샘플이 해당 지침을 준수하는 정도에 따라 라벨링하도록 요청합니다. 이러한 대화와 턴 수준 라벨은 보상 모델을 훈련하는데 사용되며, 이 모델은 이후 RLHF 과정에서 정책 모델의 샘플을 평가하는 데 활용됩니다. SFT가 교육적 지침을 어느 정도 개선하는 것으로 보이지만, 선호도 판단은 긴 대화 맥락에서 지침이 해석되고 이 따르는 방식에 미묘한 차이가 포함되는 경우가 많아 RL이 훨씬 더 효과적입니다.

2.3. 공동 훈련의 이점

교육적 행동은 대화형 AI의 일반적인 행동과 종종 상충됩니다. 이는 학습이 단순히 정보를 전달하는 과정이 아니라 발견의 과정이기 때문입니다. 우리의 지시 따르기 접근법은 특정 시스템 지시에 따라 교육 모델의 응답을 조건화함으로써 교육적 대화 데이터와 더 일반적인 상호작용을 포함한 데이터를 혼합할 수 있게 합니다. Gemini의 사후 훈련 혼합 모델과 병행 훈련함으로써, 모델이 핵심 추론 능력, 다중 모달 이해력, 사실성, 안전성, 다중 회화 특성을 '잊지' 않으면서도 새로운 유형의 지시 따르기 능력을 학습할 수 있게 합니다. 향후 훈련 방식이 진화함에 따라() LearnLM을 Gemini와 더 쉽게 동기화할 수 있을 것입니다.

3. 전문가 평가 설계

초기 기술 보고서에서 우리는 교수법 평가 설계의 분류 체계를 논의하고 서로 다른 방법론을 적용한 네 가지 인간 평가 결과를 보고했습니다(Jurenka 등 [1]의 4장 및 5장 참조). 이 분류 체계 내에서, 본 연구에서는 시나리오 기반 대화 수준 교수법 평가와 병렬 비교에 초점을 맞춥니다. 이번 새로운 평가를 위해 학습 시나리오의 명확성과 포괄성을 개선하고, 각 시나리오별 시스템 지침을 추가했으며, 교수법 평가 기준과 질문을 업데이트했습니다. 다중 대화 환경에서는 시나리오를 통해 참가자 대화를 유도하는 것이 특히 중요하다[10]. 시나리오가 없으면 인간-AI 상호작용의 제약 없는 특성으로 인해 대화가 산만해지기 쉬워 비교의 근거가 취약해진다. 반면 시나리오 기반 접근법은 서로 다른 대화형 AI 시스템의 능력을 비교할 때 상대적으로 반복 가능하고 통제된 비교를 지원한다. 시나리오 프레임워크는 평가 범위를 확장하는 데도 도움이 되어 다양한 사용 사례를 테스트할 수 있도록 보장한다.

우리의 평가 과정은 [위 그림 1](#)에 표시된 대로 세 단계로 진행됩니다. 첫째, 생태학적으로 대표적인 학습 사용 사례 분포를 식별하고 49개의 평가 시나리오 데이터베이스를 구축했습니다([3.1절](#)). 둘째, 이러한 시나리오를 바탕으로 AI 시스템과 학습자 역할을 수행하는 교육자 및 교수법 전문가 풀($N = 186$ 명) 간의 상호작용을 학습 목표, 과목, 학습 자료, 학습자 페르소나에 걸쳐 구현했습니다([3.2절](#)). 셋째, 이러한 상호작용에서 나타난 교수법의 질을 평가하기 위해 별도로 $N = 248$ 명의 교육자 및 교수법 전문가 풀을 모집하여 시스템의 성능을 검토하게 했습니다([3.3절](#)). 이 과정을 통해 시스템의 역량과 행동 양상을 이해하는 데 도움이 되는 풍부한 정량적·정성적 데이터가 생성되었습니다([3.4절](#)).

우리는 연구 목적에 대해 투명하게 소통하고, 사전 동의를 수집하며, 참여에 대해 공정하게 보상하는 등 연구 윤리 분야의 모범 사례를 따르도록 노력하고 있습니다([11](#)). 본 연구 프로토콜은 Google DeepMind 인간 행동 연구 윤리 위원회(#23 011)로부터 긍정적 의견을 받으며 독립적인 윤리 심사를 거쳤습니다.

3.1. 시나리오 설계

평가 시나리오는 대화형 AI 시스템에 대한 일관된 다중 대화 평가를 지원하는 구조화된 템플릿입니다. 시나리오는 개인과 AI 시스템 간의 상호작용에 관한 특정 '핵심 속성'을 명시합니다. 여기에는 개인의 목표, 특성, 행동 및 관련 대화 맥락이 포함됩니다. 저희가 선별한 시나리오들은 인간 참가자들에게 학문 분야, 학습 목표, 교수법에 따라 다양한 학습 맥락에서 서로 다른 유형의 학습자(예: 교실 내 학생 또는 독립적인 에드테크 사용자) 역할을 수행하도록 요청합니다. 교육 생태계의 의견과 교수법 전문가의 지원을 바탕으로 체계적인 절차를 통해 학습 시나리오 데이터베이스를 개발했습니다:

1단계: 사용 사례 도출. 시나리오 뱅크 개발을 시작하기 위해, 우리는 튜터링 및 교육에 생성형 AI를 적용하려는 에드테크 기업, 교육 기관 및 Google 제품 팀으로부터 피드백을 수집했습니다. 실제 교육 환경에서 생성형 AI에 대해 그들이 인식한 일반적인 사용 사례, 프롬프트, 기회 및 과제를 공유해 달라고 요청했습니다. 우리는 평가 접근 방식에 반영해야 할 공통 주제를 식별하기 위해 팀 차원에서 이 피드백을 수집하고 분석했습니다.

2단계: 시나리오 설계. 이러한 사용 사례, 기회 및 과제를 바탕으로 구조화된 시나리오 템플릿(부록 B.1의 "시나리오 구조 및 내용" 참조)과 시나리오 생성을 주도하기 위한 구체적인 프로토콜을 초안 작성하였습니다. 여기에는 각 속성에 대한 일련의 안내 질문이 포함됩니다(부록 B.2의 "시나리오 생성 프로토콜" 참조).

3단계: 시나리오 생성 및 정교화. 다음으로 우리는 협업적이며 반복적인 과정을 통해 시나리오 데이터베이스를 구축했습니다. 우리 팀원들(학생 교육 및 교사 연수 분야에서 다년간의 전문 경험을 가진 두 명 포함)은 2단계에서 마련한 템플릿과 가이드 질문을 활용하여 각자 시나리오 초안을 작성했습니다. 이후 시나리오 초안을 공동으로 검토하며 각 시나리오의 명확성, 완전성, 정확성, 그리고 1단계에서 정의한 교육 원칙 및 사용 사례와의 관련성을 평가했습니다. 다양한 학습 목표, 페르소나, 과목 영역에 걸친 시나리오의 전반적인 분포를 가중치 부여하여 분석하고, 추가 개발이 필요한 부분을 표시했습니다.

이 과정을 통해 핵심 학문 분야 전반에 걸쳐 49개의 다양한 시나리오 데이터베이스가 구축되었습니다(예시 참조: 부록 B.3). 이러한 기초 데이터베이스 구축 외에도, 이후 전문 교육 분야 (구체적으로는 의학 교육; 부록 C 참조)에서 실행 가능성 연구를 수행하여 이 절차의 견고성과 재현성을 검증하였습니다.

3.2. 대화 수집

두 번째 단계에서는 평가 시나리오에 명시된 대로 인간 참가자들이 학습자 역할을 맡아 AI 시스템과 상호작용하는 대화 코퍼스를 수집했습니다. 교육 시나리오에서 학습자 행동을 효과적으로 시뮬레이션하기 위해, 고급 학위를 보유하고 2년 이상의 튜터 경력을 가진 교육학 전문가 풀 $N = 168$ 명을 모집했습니다.

대화 수집 세션은 시나리오 역할극에 대한 간단한 훈련으로 시작되었습니다(그림 2, 단계 1 참조). 훈련 종료 시 퀴즈를 통과한 참가자는 수행할 시나리오를 선택했습니다(그림 2, 단계 2 참조). 대화 수집은 쌍을 이루어 진행되었으며, 동일한 참가자가 먼저 한 AI 시스템과 시나리오를 수행한 후 다른 시스템과 수행했습니다. 각 쌍에는 LearnLM과 비교 대상 시스템이 포함되었습니다. 각 대화 쌍에 대해 시스템 순서를 무작위로 배정했으며 참가자에게 시스템 라벨을 제공하지 않았습니다. 각 대화 쌍 내에서 모델들은

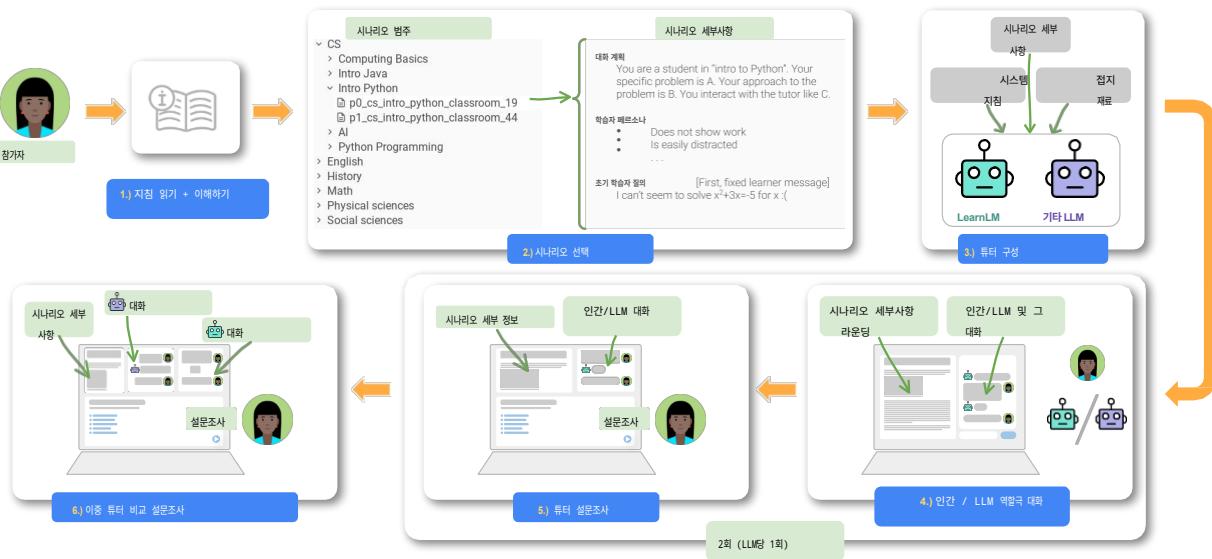


그림 2 | 교육 시나리오 기반 대화 생성 워크플로우. 참가자는 시나리오에 정의된 대로 프롬프트 모델과 대화를 수행합니다. 이후 참가자는 모델 간 품질 및 선호도를 기록하는 설문지를 작성합니다.

시나리오에 명시된 대로 동일한 시스템 지침, 기초 자료 및 초기 학습자 질의를 컨텍스트로 사용했습니다(그림 2, 단계 3 참조). 시스템 API에서 요구하는 사소한 사양 차이점을 제외하고 모든 입력을 동일하게 형식화했습니다.

템플릿에 명시된 대로, 각 시나리오에는 학습자를 위한 초기 질의가 포함되었으며, 이는 참가자를 대신하여 대화를 시작하기 위해 자동으로 전송되었습니다. AI 시스템이 해당 질의에 응답한 후, 참가자는 시나리오에 제공된 정보에 따라 상호작용을 계속했습니다. 참가자는 상호작용을 종료하기 전에 최소 10회 이상의 대화 턴(즉, 학습자 5턴과 튜터 5턴)을 계속해야 했습니다.

각 대화 종료 후(그림 2, 단계 4 참조), 참가자들은 시스템과의 상호작용 경험을 공유하기 위한 간단한 설문지를 작성했습니다(그림 2, 단계 5 및 부록 B.4 참조). 또한, 두 차례의 대화 쌍이 끝난 후 참가자들은 두 시스템에 대한 인상을 비교하는 데 초점을 맞춘 또 다른 설문지를 작성했습니다(그림 2, 단계 6 및 부록 B.5 참조). 참가자들은 이후 에서 새로운 시나리오를 선택하여 두 차례의 추가 대화를 시작하거나 세션을 종료할 수 있었습니다.

3.3. 교육학적 평가

마지막으로 세 번째 단계에서는 또 다른 교육 전문가 집단($N = 228$ 명)을 모집했습니다. 이들은 다시 한번 고급 학위와 2년 이상의 튜터 경력을 보유한 전문가들로, 이 대화들을 분석하고 다양한 AI 모델의 교육적 역량을 평가했습니다.

각 평가 세션은 평가 목표와 시나리오 템플릿에 대한 간단한 교육으로 시작되었습니다. 참가자마다 검토할 시나리오를 무작위로 배정했습니다. 검토 후에는 해당 시나리오에서 평가할 대화 쌍을 무작위로 배정했습니다(즉, 대화 수집 단계에서 단일 참가자로부터 수집된 대화 쌍). 참가자들은 한 번에 하나의 대화 기록을 검토했습니다. 기록을 읽은 후 참가자들은 해당 대화에서 AI 시스템의 교육적 성능에 초점을 맞춘 설문지에 답변했습니다(부록 B.6 참조). 대화 쌍마다 참가자들은 두 시스템에 대한 자신의 평가를 비교하는 추가 간단한 설문지를 작성했습니다(부록 B.7 참조). 평가자 간 변동성의 영향을 줄이기 위해 각 대화 쌍에 대해 세 개의 독립적인

평가를 수집하여 평가자 간 변동성의 영향을 줄이려 했습니다.

3.4. 분석

우리는 정량적 분석을 위해 베이지안 통계적 프레임워크를 활용합니다. 가설의 확률을 직접 정량화하고 명확하며 해석 가능한 불확실성 척도를 제공함으로써, 베이지안 분석은 실제 세계에 배포될 AI 시스템을 평가하기 위한 실용적이고 유익한 접근법을 제시합니다.

본 연구 설계는 참가자들의 반복 측정을 포함합니다. 즉, 학습자 역할을 맡은 각 참가자는 각 시스템과 여러 차례 상호작용했으며, 각 전문가는 각 시스템을 여러 차례 평가했습니다. 이러한 비독립성을 고려하고 추정값에 대한 확신을 인위적으로 부풀리지 않기 위해, 우리는 계층적 회귀 분석[12]을 통해 데이터를 분석합니다. 부록 B.8에서는 통계적 방법론을 보다 상세히 설명합니다.

또한, 두 시스템으로 각 시나리오를 역할극한 후 전문가들로부터 수집한 개방형 의견과 피드백에 대한 질적 분석을 수행했습니다(2단계)³. 특히 참가자들의 자유형 응답에서 학습자-시스템 상호작용과 관련된 일반적인 주제를 식별하고 정제했습니다. 그런 다음 각 주제의 유무를 기준으로 개별 응답을 코딩했습니다. 주석 작업에 편향을 방지하기 위해 이 과정에서 시스템의 신원을 익명화했습니다. 부록 B.9에는 분석을 통해 개발한 코드북이 제시되어 있습니다.

4. 결과

LearnLM을 동시대의 주요 제품들(2024년 10월 1일 기준), 특히 GPT-4o⁴, Claude 3.5 Sonnet⁵, Gemini 1.5 Pro⁽⁶⁾와 비교했습니다. 이 특정 평가를 수행한 이후, 이들 모델은 각각 업데이트되어 새 버전이 출시되었습니다. 따라서 본 결과는 특정 시점의 비교로 이해되어야 하며, 이는 우리 접근법의 효과성을 평가하고 교육 분야에 대한 지속적인 투자의 기준점을 마련하기 위한 것입니다.

총 2360건의 대화 세트를 수집했으며, 이는 총 58,459개의 학습자 및 모델 메시지로 구성됩니다. 해당 대화들에 대해 10,192건의 전문가 평가를 수집했으며, 평균적으로 각 대화 쌍마다 세 명의 전문가가 검토했습니다. 그림 3은 평가 대상 시스템들이 수집된 대화 전반에 걸쳐 응답 길이 분포에서 현저한 차이를 보임을 보여줍니다. 이는 Gemini 1.5 Pro와 LearnLM 간 차이도 포함됩니다. 종합적으로 볼 때, 길이(길이)와 인지된 품질 간에는 명확한 상관관계가 관찰되지 않습니다([13] 참조).

분석을 시작하며 전문가 평가자들의 교육적 평가와 선호도 점수를 검토합니다. 이후 모델과 상호작용하는 학습자 역할을 수행한 참가자들의 직접 피드백을 탐구합니다. 즉, 2단계에서 수집한 상호작용 데이터로 회귀하기 전에 3단계의 교육적 발견 사항을 제시합니다. 이 분석에서 몇 가지 명확한 패턴이 도출됩니다.

첫째, 비교 선호도 평가(그림 4)는 다섯 가지 비교 평가 범주 모두에서 GPT-4o보다 LearnLM에 대한 강한 선호도를 보여줍니다. 전문가들은 전반적인 교수법("어느 튜터가 더 나은 튜터링을 보여주었는가?")에서 LearnLM에 대한 가장 강한 선호도를 표명했습니다. 또한 그들은

³ 질적 분석을 위해 우리는 각 튜터와의 학습 경험에 대한 직접적인 통찰력을 제공하는 2단계 피드백을 우선적으로 고려했습니다.

⁴ GPT-4o 버전 2024-08-06, <https://platform.openai.com/docs/models/gpt-4o>.

⁵ Claude 3.5 Sonnet 버전 2024-06-20, <https://docs.anthropic.com/en/docs/about-claude/models>.

⁶ 2024-09-24의 Gemini 1.5 Pro-002, <https://cloud.google.com/vertex-ai/docs/learn/model-versions>.

시스템	버전	평균 대화당 턴 수	평균 발언당 단어 수
LearnLM	2024-11-19	11.0	174
Gemini 1.5 Pro	2024-09-24	10.3	130
GPT-4o	2024-08-06	10.1	137
Claude 3.5 소네트	2024-06-20	9.7	179

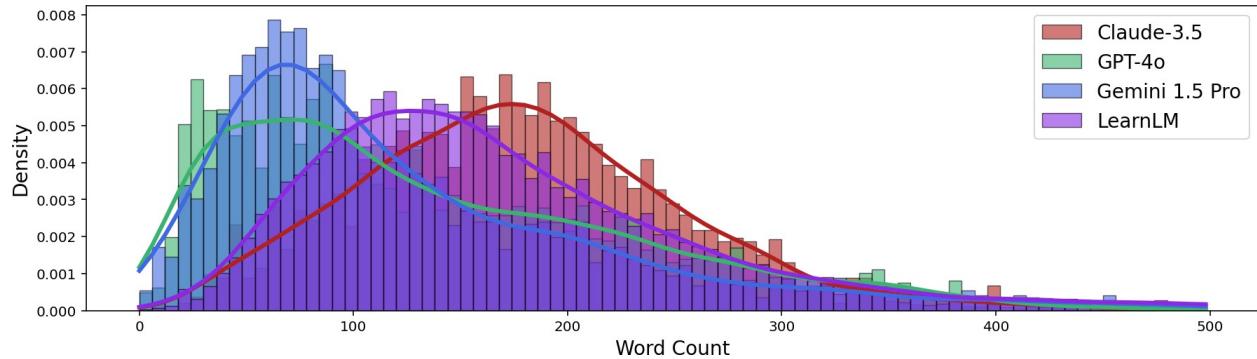


그림 3 | (위) 비교 대상 특정 대규모 언어 모델(LLMs) 및 수집된 모든 대화의 집계 통계: 대화당 평균 턴 수 및 턴당 평균 단어 수; (아래) 각 모델별 턴당 사용 단어 수의 히스토그램.

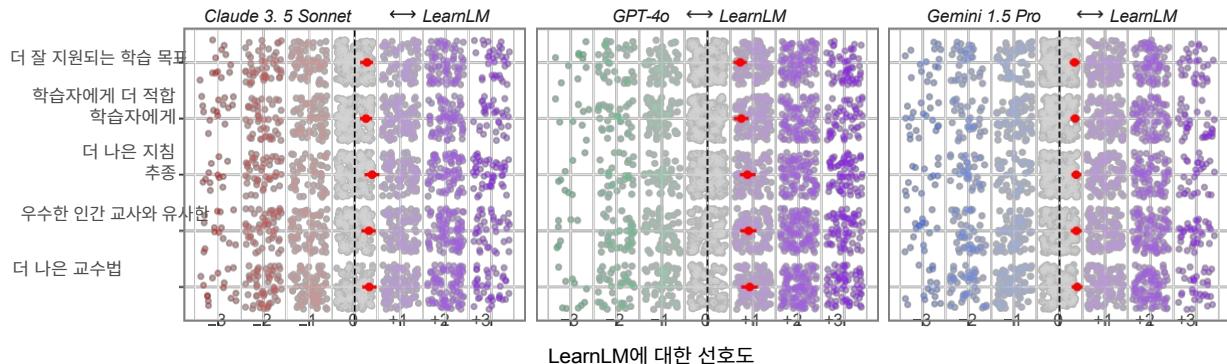


그림 4 | LearnLM 및 동시대 시스템(Claude 3.5 Sonnet, GPT-4o, Gemini 1.5 Pro)에 대한 교육 전문가들의 선호도. 산점도는 7점 척도 선호도 평가의 기본 분포를 나타냅니다. 수집한 평가 수가 방대하므로, 각 척도당 500개 평가로 비례 다운샘플링했으며, 선호도 척도에 따라 색상 코딩(진한 보라색은 LearnLM에 대한 강한 선호도)하고 가독성을 위해 각 척도값 주변에 무작위 진동을 적용했습니다. 빨간색 점과 오차 막대는 각 척도에 대한 추정 평균과 95% 신뢰 구간을 나타냅니다.

LearnLM이 Claude 3.5 Sonnet 및 Gemini 1.5 Pro에 비해 유사하지만 더 낮은 선호도를 보였습니다. LearnLM 훈련에 Gemini 1.5 Pro를 적용했기 때문에, 두 모델 간 비교는 교육적 데이터 추가로 인한 변화를 직접 반영합니다([2절 참조](#)).

둘째, 그림 5는 각 모델의 교육학적 평가 기준에 대한 평균 성능을 보여줍니다. 전문가들은 개별 교육학적 특성을 7점 척도로 평가했습니다. 평균적으로 각 시스템은 이 평가에서 모든 평가 항목 범주에 걸쳐 긍정적인 평가를 받았습니다. 전문가들은 LearnLM에 모든 평가 항목 범주와 거의 모든 29개 평가 질문에서 최고 점수를 부여했으며, 특히 능동적 학습 유도, 메타인지 심화, 호기심 자극 항목에서 큰 차이로 앞섰습니다.

셋째, 그림 6은 각 시스템이 참가자의 튜터링 주제 관심도, 향후 모델 사용 의향[14], 모델의 역량 및 친근감 인식[15, 16]을 얼마나 증진시켰는지를 나타냅니다. 참가자들은 상대적으로 유사한



그림 5 | 7점 응답 척도("전혀 동의하지 않음"부터 "매우 동의함")로 본 교수법 평가표 각 범주별 시스템 평가. 오차 막대는 평균에 대한 사후 분포의 95% 신뢰 구간을 반영함.

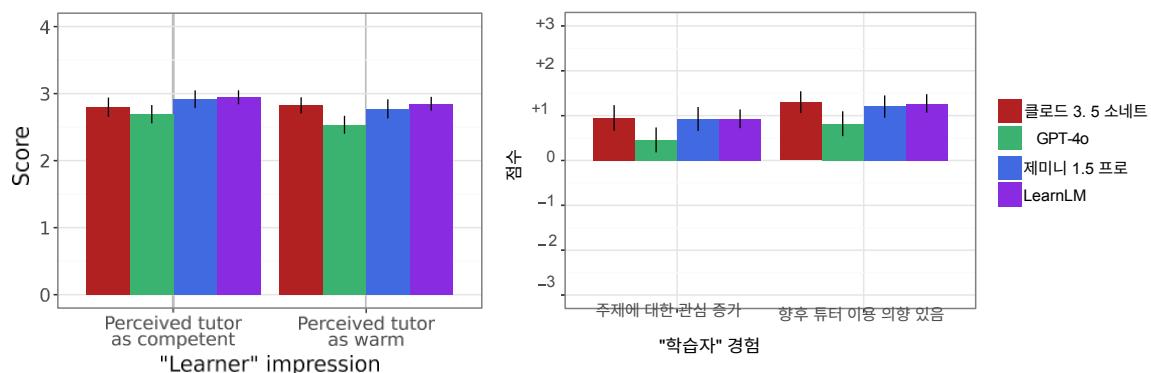


그림 6 | 교육 시나리오에서 학습자 역할을 수행한 교육 전문가들이 공유한 인상. 오차 막대는 평균에 대한 사후 분포의 95% 신뢰 구간을 반영합니다. 참가자들은 인상 항목에 대해 5점 척도("전혀 그렇지 않다"부터 "매우 그렇다")로, 경험 항목에 대해 7점 척도("전혀 동의하지 않는다"부터 "매우 동의한다")로 응답했습니다.

LearnLM, Gemini 1.5 Pro, Claude 3.5 Sonnet에 대한 경험. 반면 참가자들은 GPT-4o가 자신의 관심도에 미치는 영향, 인식된 친근함, 인식된 유용성 측면에서 더 약한 경험을 나타냈다. 역할극을 수행한 전문가들이 학생들을 완벽하게 대변하지는 않지만, 그들의 인상은 AI 튜터링 상호작용의 사용자 경험에 대한 예비적 통찰을 제공하는 데 도움이 된다.

넷째, 역할극 학습자 선호도 주제 분석을 위해 수집한 1024개 설명 중 무작위로 203개(약 20%)를 하위 표본 추출했습니다(테마 별 예시 발췌문 포함 자세한 내용은 표 1 참조). 하위 표본에서 가장 일관되게 나타난 주제는 `is_engaging`(하위 표본 설명 72개에 등장)이었습니다.

`conversation_style` (67개 설명), 그리고 `gives_away_answers` (50개 설명).

참가자들이 LearnLM을 다른 모델보다 선호한다고 보고했을 때, 그들의 설명에는 주제 `keeps_on_topic`, `challenges_learner`, `gives_away_answers` 가 포함될 가능성이 더 높았습니다. 반면 LearnLM보다 다른 모델을 선호한 참가자들은 설명에서 `clarity`, `info_amount`, `conversation_style` 주제들을 다루는 경향이 있었습니다. 전반적으로 전문가 커뮤니티는 LearnLM이 단순히 답을 알려주기보다는 주제 유지와 학습자의 개념에 대한 탄탄한 이해 유도 측면에서 더 우수하다고 평가하는 경향이 있었습니다. 반면, 이들 전문가는 LearnLM이 정보 전달 방식이나 대화 스타일 측면에서 때때로 덜 적합하다고 판단하기도 했습니다.

주제	참가자가 LearnLM 을 선호한 횟수 (94건 중)	참가자가 다른 모델 을 선호한 횟수 (80건 중)	예시 응답
주제를 유지합니다	20 (21.2%)	8 (10%)	“[LearnLM]은 제가 주의를 분산시키는 것을 용납하지 않았습니다” “[LearnLM]은 훨씬 더 효과적으로 주제를 유지할 수 있었습니다” “[다른 튜터]도 제가 다시 집중하도록 만드는데 훨씬 더 잘했어요” “제대로 집중하게 하는 데 훨씬 더 능숙했습니다”
학습자_도전과제	31 (33.0%)	13 (16.3%)	“분명히 [LearnLM]이 더 좋았습니다 [...] [다른 튜터]는 제가 잘하도록 독려하지 않았습니다” “[LearnLM]이 제가 성장하고 배울 수 있도록 노력하는 것 같았어요, 그냥 제 말에 동의만 하는 게 아니라” “[다른 튜터]는 흥미로운 질문을 해서 더 깊이 생각하게 했어요”
정답을 알려줌	32 (34.0%)	15 (18.8%)	“[LearnLM]은 정말로 질문에 답하는 단계에 저를 몰입시켰습니다” 반면 [다른 튜터]는 그냥 정답만 알려줬어요” “[LearnLM]은 답을 주는 것보다 답을 도출하는 과정에 더 집중했어요” “[LearnLM]은 학생이 분명히 필요로 할 때조차 도움을 주기에는 자나치게 소극적이었다”
명확성	15 (16.0%)	16 (20.0%)	“[다른 튜터에 대한] 지원 구조는 학생이 따라가기 좀 더 명확했다” “[다른 튜터는] 더 작고 단순한 것부터 시작했어요” “[LearnLM의] 답변이 더 명확하다고 생각했어요”
info_amount	19 (20.2%)	20 (25.0%)	“[다른 튜터는] [...] 더 간결했어요” “[다른 튜터]는 제가 필요한 모든 것을 제공해 주셨다” “[LearnLM]은 이 '복잡한' 주제를 더 이해하기 쉬운 조각들로 나누는데 더 잘 해냈습니다” “주제를 더 소화하기 쉬운 조각으로 나누는 데 더 잘 해냈습니다”
conversation_style	30 (31.9%)	29 (36.3%)	“저는 [...] [LearnLM]이 약간 생색내는 느낌이 들었습니다” “[다른 튜터는] 더 따뜻하고 친근하게 느껴졌다” “[LearnLM]은 더 따뜻하고 격려적이었다”

표 1 | LearnLM 선호(상단 3행) 또는 다른 모델 선호(하단 3행)에 대한 학습자 선호도 설명에서 더 자주 나타난 주제. 이 표는 (i) 전체 선호도 설명의 최소 10% 이상이 언급한 주제와 (ii) LearnLM 선호 설명과 다른 모델 선호 설명 간 발생 비율이 극단적으로 나타난 주제를 보여줍니다.

4.1. 안전성 평가

초기 기술 보고서[1] 및 Gemini 기술 보고서[3, 17]에 기술된 과정과 유사하게, Google DeepMind의 책임 있는 개발 및 혁신 팀과 Google의 신뢰 및 안전 팀과 협력하여 LearnLM에 대한 안전성, 책임성 및 보증 평가를 수행했습니다. 이는 Gemini의 모델 정책과 학습 특화 모델 정책 준수를 보장하기 위함이었습니다.

모델 카드 업데이트된 접근 방식은 Gemini를 따른 교육적 지침 및 공동 훈련에 중점을 둡니다. 따라서 훈련 및 안전 평가 절차는 이제 Gemini 1.5와 완전히 일치합니다. 해당 모델 카드는 보고서[3]의 부록 12, 표 45를 참조하십시오. 참고로, 학습 전용 데이터셋 큐레이션을 포함한 모델링 접근법은 2절에서 설명합니다. 초기 기술 보고서[1]에는 LearnLM의 원본 모델 카드와 함께 이 연구 분야의 윤리적 위험 및 한계에 대한 광범위한 논의가 수록되어 있습니다.

5. 결론

우리는 학습 활용 사례를 위한 기초 모델 개선 동기와 접근법을 설명했으며, 이는 원하는 행동을 조건화하기 위해 시스템 지침에 의존합니다. Gemini의 사후 훈련 혼합 모델을 업데이트하여 시연 데이터(SFT를 통해)와 인간 선호도 데이터(보상 모델 및 RLHF를 통해)를 추가함으로써 모델이 다양한 교육적 지침을 따르도록 가르쳤습니다. 이후 결과물인 LearnLM 모델을 유사 모델들과 비교 평가한 결과, 특히 지시 따르기 능력에서 LearnLM에 대한 상당한 선호도가 확인되었으며, 더 넓게는 다양한 교육적 차원에서도 우위를 보였습니다. 본 연구는 LearnLM의 성과를 Gemini⁽⁷⁾에 적용하여 학습 활용 사례를 위한 Gemini 개선 노력의 시작을 의미합니다. 교육적 지시 따르기 기능을 지속적으로 개선하여 교사 및 교육 제품 개발자의 편의성을 위해 교육적 행동 지정이 최대한 간단하고 직관적으로 이루어지도록 할 것입니다.

모델 개선 외에도 평가 방법론에 대한 추가 업데이트를 계획 중입니다. 첫째, AI 시스템의 교육적 평가를 위한 보편적 프레임워크에 대한 합의 도출을 목표로 합니다. 현재 교육 평가 기준(부록 B.7 참조)은 학습 과학 원리를 기반으로 하지만, 모든 학습자에게 적합하고 광범위한 교육계의 신뢰와 승인을 얻기 위해 다양한 이해관계자들과의 긴밀한 협력이 필요합니다.

둘째, 우리는 사전에 정의된 교육학적 기준에 따라 모델의 성능을 측정하는 내재적 평가에서 벗어나 학습 성과와 같은 영향을 측정하는 외재적 평가로 전환하고자 합니다. 내재적 평가는 실행 속도가 빠르고 모델의 결함을 직접적으로 식별할 수 있어 모델 개발에 유용합니다. 그러나 능동적 학습 장려 및 인지 부하 관리와 같은 우리 평가 기준의 핵심 원칙들은 광범위하게 합의되고 증거 기반임에도 불구하고[18], 그 결과가 학습 성과 개선으로 얼마나 효과적으로 연결되는지는 불분명합니다. 이 분야가 성숙해지고 AI 시스템이 튜터링 대화의 기본을 숙달함에 따라 외재적 평가가 더 중요한 역할을 할 가능성이 높습니다. 최근에는 학습 성과 개선을 입증하는 데[19, 20]와 서로 다른 시스템 및 프롬프트를 비교하는 데[21] 모두 활용되고 있습니다.

마지막으로, 우리는 핵심 학문 과목을 넘어선 평가 영역을 탐구하기 시작했습니다. 의학 교육 분야에서의 초기 타당성 연구(부록 C)는 우리의 접근법이 전문 분야로 효과적으로 확장될 수 있음을 확인시켜 줍니다. 다양한 교육 환경에서 사용하기 위해 Gemini를 지속적으로 개선해 나가는 가운데, LearnLM 적용 사례로부터 얻은 통찰력을 환영하며, 이를 통해 교육 및 학습 분야에서 AI의 잠재력을 실현하는 데 기여하고자 합니다[22–24].

참고문헌

- [1] Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, et al. 교육용 생성형 AI의 책임감 있는 개발을 향하여: 평가 중심 접근법. *arXiv 사전 인쇄본 arXiv:2407.12687*, 2024.
- [2] 다니엘 M 지글러, 니산 스티엔논, 제프리 우, 톰 B 브라운, 알렉 래드포드, 다리오 아모데이, 폴 크리스티아노, 제프리 어빙. 인간 선호도를 통한 언어 모델 미세 조정. *arXiv 사전 인쇄본 arXiv:1909.08593*, 2019.
- [3] Gemini 팀, 페트코 게오르기예프, 빙 이안 레이, 라이언 버넬, 리빈 바이, 안몰 굴라티, 개럿 탄저, 데미엔 빈센트, 주평 판, 시보 왕 외. Gemini 1.5: 수백만 토큰의 컨텍스트를 아우르는 다중 모달 이해의 해방. *arXiv 사전 인쇄본 arXiv:2403.05530*, 2024.

⁷ 출판 시점에, 당사의 일부 데이터는 이미 Gemini 2 모델[4]에 추가되었습니다.

- [4] Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 에이전트 시대를 위한 새로운 AI 모델, Gemini 2.0 소개. <https://blog.google/technology/google-deepmind/> google-gemini-ai-update-december-2024/, 2024.
- [5] 롱 우양, 제프리 우, 쉬 장, 디오고 알메이다, 캐를 웨인라이트, 파멜라 미쉬킨, 종 장, 산디니 아가르왈, 카타리나 슬라마, 알렉스 레이 외. 인간 피드백을 통한 지시사항 수행을 위한 언어 모델 훈련. *신경정보처리시스템 발전*, 35: 27730–27744, 2022.
- [6] Ethan Mollick, Lilach Mollick. AI 할당하기: 학생들을 위한 일곱 가지 접근법과 프롬프트. *arXiv 사전 인쇄본 arXiv:2306.10052*, 2023.
- [7] Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu 외. 다중 제약 조건 조합을 통한 복잡한 지시 따르기 벤치마킹. *arXiv 사전 인쇄본 arXiv:2407.03978*, 2024.
- [8] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 대규모 언어 모델을 위한 명령어 추종 평가. *arXiv 사전 인쇄본 arXiv:2311.07911*, 2023.
- [9] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, Dong Yu. Infobench: 대규모 언어 모델의 지시 따르기 능력 평가. *arXiv 사전 인쇄본 arXiv:2401.03601*, 2024.
- [10] 루자인 이브라힘, 사프론 황, 라마 아마드, 마르쿠스 안데를룽. 정직 AI 평가를 넘어: 대규모 언어 모델의 위험 및 유해성에 대한 인간 상호작용 평가 발전. *arXiv 사전 인쇄본 arXiv:2405.10632*, 2024.
- [11] Kevin R McKee. 인공지능 연구에서의 인간 참여자: 실천에서의 윤리와 투명성. *IEEE Transactions on Technology and Society*, 5(3):279–288, 2024. doi: 10.1109/TTS.2024.3446183.
- [12] 앤드류 젤만, 존 B. 칼린, 할 S. 스텐, 도널드 B. 루빈. 『베이지안 데이터 분석』. 채프먼 앤드 홀/CRC, 1995.
- [13] Wei-Lin Chiang Tianle Li, Anastasios Angelopoulos. 스타일이 중요한가? 챗봇 분야에서 스타일과 실질을 분리하기. <https://blog.lmarena.ai/blog/2024/style-control/>, 2024년 8월.
- [14] 프레드 D 데이비스. 인지된 유용성, 인지된 사용 용이성 및 정보 기술의 사용자 수용. *MIS 큐터리*, 1989.
- [15] Susan T Fiske, Amy JC Cuddy, Peter Glick. 사회적 인지의 보편적 차원: 친근감과 유능함. *Trends in cognitive sciences*, 11(2):77–83, 2007.
- [16] 케빈 R. 맥키, 바이 슈춘즈, 수잔 T. 피스크. 인간은 인공 지능에서 친근감과 유능함을 인지한다. *iScience*, 26(8), 2023. doi: 10.1016/j.isci.2023.107256.
- [17] 제미니 팀, 로한 아닐, 세바스티앙 보르고, 우용휘, 장-밥티스트 알라이라크, 유자휘, 라두 소리컷, 요한 살크워크, 앤드류 M. 다이, 안야 하우트 외. 제미니: 고성능 다중 모달 모델 군. *arXiv 사전 인쇄본 arXiv:2312.11805*, 2023.
- [18] Paul A Kirschner and Carl Hendrick. *학습이 일어나는 방식: 교육 심리학의 선구적 연구와 실제적 의미*. Routledge, 2020.

- [19] Gregory Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, Gregorio Ponti. AI 튜터링이 능동적 학습을 능가한다. 2024.
- [20] Rose E Wang, Ana T Ribeiro, Carly D Robinson, Susanna Loeb, Dora Demszky. 튜터 코파일럿: 실시간 전문성 확장을 위한 인간-AI 접근법. *arXiv* 사전 인쇄본 *arXiv:2410.03017*, 2024.
- [21] 함사 바스타니, 오스버트 바스타니, 알프 순구, 하오센 게, 오즈게 카박치, 레이 마리안. 생성형 AI는 학습에 해로울 수 있다. *SSRN*에서 이용 가능, 4895486, 2024.
- [22] 전국교육협회(National Education Association, NEA). *Teaching and learning in the age of artificial intelligence.* <https://www.nea.org/resource-library/artificial-intelligence-education/div-teaching-and-learning-age-artificial-intelligence>. 접속일: 2024-12-10.
- [23] 킴벌리 로미스, 파멜라 제프리스, 앤서니 팔라타, 멜라니 세이지, 자비드 셰이크, 칼 세페리스, 앤리슨 훨런. 보건 전문직 교육자를 위한 인공지능. *NAM 관점*, 2021, 2021.
- [24] 산제이 V 데사이, 제시 버크-라펠, 킴벌리 D 로미스, 켈리 카버자지, 주디 리처드슨, 셀리아 레어드 오브라이언, 존 앤드루스, 케빈 헥만, 데이비드 헨더슨, 찰스 G 프로버 외. 정밀 교육: 의학 분야 평생 학습의 미래. *학술 의학*, 10–1097쪽, 2023.
- [25] 케빈 R. 맥키, 바이 슈춘즈, 수잔 T. 피스크. 인간-에이전트 협력에서의 친근감과 유능함. *자율 에이전트 및 다중 에이전트 시스템*, 38(1):23, 2024.

기여 및 감사의 글

핵심 기여자 아비닛 모디, 아디티야 스리칸트 비루보틀라, 알리야 리스벡, 안드레아 후버, 브렛 월트셔, 브라이언 베프레크, 다니엘 길릭, 다니엘 카젠버그, 데릭 아메드, 이리나 유렌카, 제임스 코한, 제니퍼 세, 줄리아 월코프스키, 카이즈 알라라키야, 케빈 R. 맥키, 리사 왕, 마르쿠스 쿠네쉬, 마이크 샤크만, 미루나 피슬라르, 니킬 조시, 파르사 마흐무디에, 폴 준, 사라 월트버거, 샤키르 모하메드, 샤샨크 아가르왈, 슈밤 밀린드 팔, 이선재, 테오필로스 스트리노풀로스, 위젠 코.

기여자 에이미 왕, 안킷 아난드, 아비슈카르 부프찬드, 댄 와일드, 디비야 판디아, 필립 바르, 가스 그레이엄, 홀거 빈네뫼러, 마비쉬 나그다, 프라티크 콜하르, 르네 슈나이더, 샤오지안 주, 스테파니 찬, 스티브 앤들로우스키, 비크네쉬 사운더라자, 앤시스 아사엘.

역할은 다음과 같이 정의됩니다: **핵심 기여자는** 본 보고서에 제시된 작업에 직접적이고 중대한 영향을 미쳤습니다. **기여자는** 본 보고서에 제시된 작업에 기여했습니다. 각 역할 내 순서는 알파벳 순이며 기여도 순위를 나타내지 않습니다.

감사의 말씀

본 작업은 LearnLM 프로젝트의 일환으로 수행되었으며, 이는 Google DeepMind(GDM), Google Research(GR), Google LearnX, Google Health, Google Creative Lab, YouTube Learning, YouTube Health 등 Google 내 여러 부서가 참여한 크로스-Google 프로젝트입니다. 본 기술 보고서는

교육적 지시 추적 개선에 초점을 맞춘 이 기술 보고서는 광범위한 노력의 일부에 불과하며, 위 기여자 목록에는 직접적인 기여만 포함됩니다.

구글 내 수많은 팀의 혁신과 노력 덕분에 저희의 작업이 가능해졌습니다. 다음 분들의 지원에 감사드립니다: 아제이 칸난, 아난드 라오, 아니샤 초두리, 에이프릴 (솔러) 마노스, 던 챈, 다르티 다미, 에드워드 그레펜스테트, 갈 엘리단, 히만슈 카텔루, 하우메 산체스 엘리아스, 자오 쑨, 조쉬 카필루토, 조티 굽타, 칼페쉬 크리슈나, 로렌 와이너, 맥 맥앨리스터, 마나 자부르, 마이클 하웰, 미리암 슈나이더, 무크타 아난다, 니르 레빈, 니브 에프론, 라이언 뮐러, 사프완 초두리, 샤암 우파디야이, 스베틀라나 그랜트, 테자시 라트카르, 윌리엄 웡, 야엘 하라마티. 또한 Google DeepMind의 Gemini 팀, Google DeepMind의 책임 있는 개발 및 혁신 팀, 책임 있는 엔지니어링 팀, 아동 안전 팀, 그리고 Google의 신뢰 및 안전 팀에 감사드립니다. 마지막으로, 이 프로젝트를 실현할 수 있도록 지원해 주신 모든 리더와 후원사 여러분께 감사의 말씀을 전합니다.

A. 추가 결과

A.1. 학습자 역할극 참가자의 선호도

학습자 역할을 맡은 참가자들은 네 가지 비교 평가 범주 모두에서 GPT-4o보다 LearnLM을 선호하는 것으로 나타났다(그림 7). 전문가들은 전반적인 교수법("어느 튜터가 더 나은 튜터링을 보여주었는가?")과 우수한 인간 튜터와의 유사성("어느 튜터가 매우 훌륭한 인간 튜터와 더 비슷했는가?")에서 LearnLM에 대한 가장 강한 선호도를 나타냈다. 이 참가자들은 LearnLM과 Gemini 1.5 Pro 사이, 또는 LearnLM과 Claude 3.5 Sonnet 사이에는 큰 선호도 차이를 보이지 않았다.

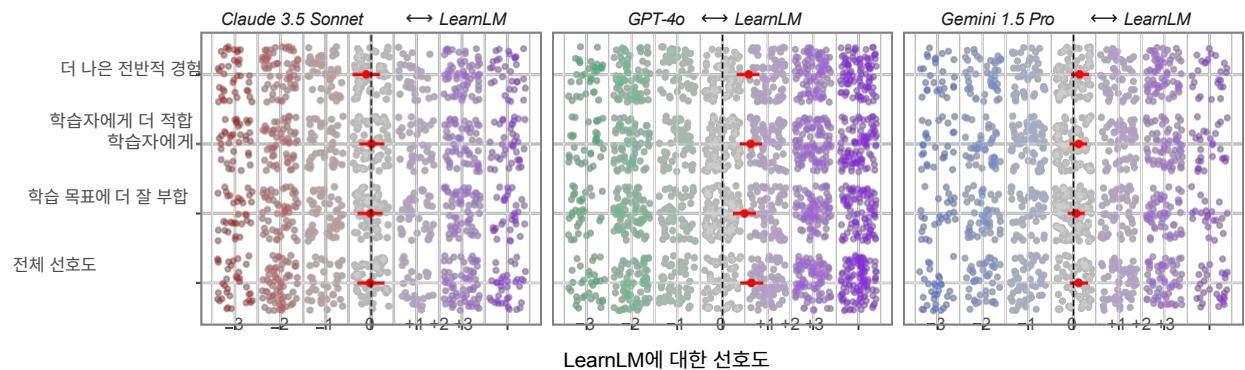


그림 7 | 학습자 역할을 맡은 교육 전문가들의 LearnLM 및 기타 동시대 모델(Claude 3.5 Sonnet, GPT-4o, Gemini 1.5 Pro)에 대한 선호도. 산점도는 7점 척도 선호도 평가의 기본 분포를 나타냅니다. 수집된 평가 수가 많기 때문에, 이 산점도는 측정 항목당 500개 평가로 비례 다운샘플링되었습니다. 빨간색 점과 오차 막대는 각 측정 항목에 대한 추정 평균과 95% 신뢰 구간을 나타냅니다.

A.2. 수집된 대화에서의 학습자 품질

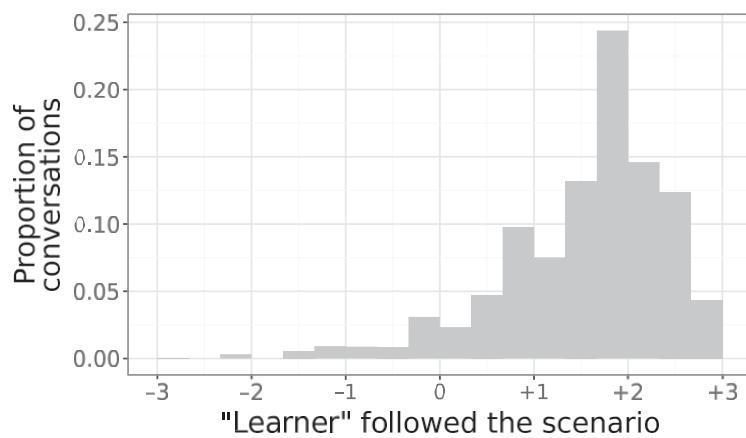


그림 8 | 교육적 평가 과정의 시작 단계에서, 우리는 전문가들에게 대화 기록에 등장하는 인간 참여자들이 시나리오 지침을 얼마나 충실히 따랐는지(즉, 시나리오에서 학습자 역할을 얼마나 효과적으로 수행했는지) 7점 척도로 평가해 달라고 요청했습니다. 이 그래프는 대화 기록별로 응답을 그룹화하고 평균화한 결과를 보여줍니다. 이러한 종합 평가는 "학습자"가 대화 기록의 93.2%에서 시나리오 지침을 따랐음을 나타냅니다.

A.3. 교육적 평가: 상세 결과

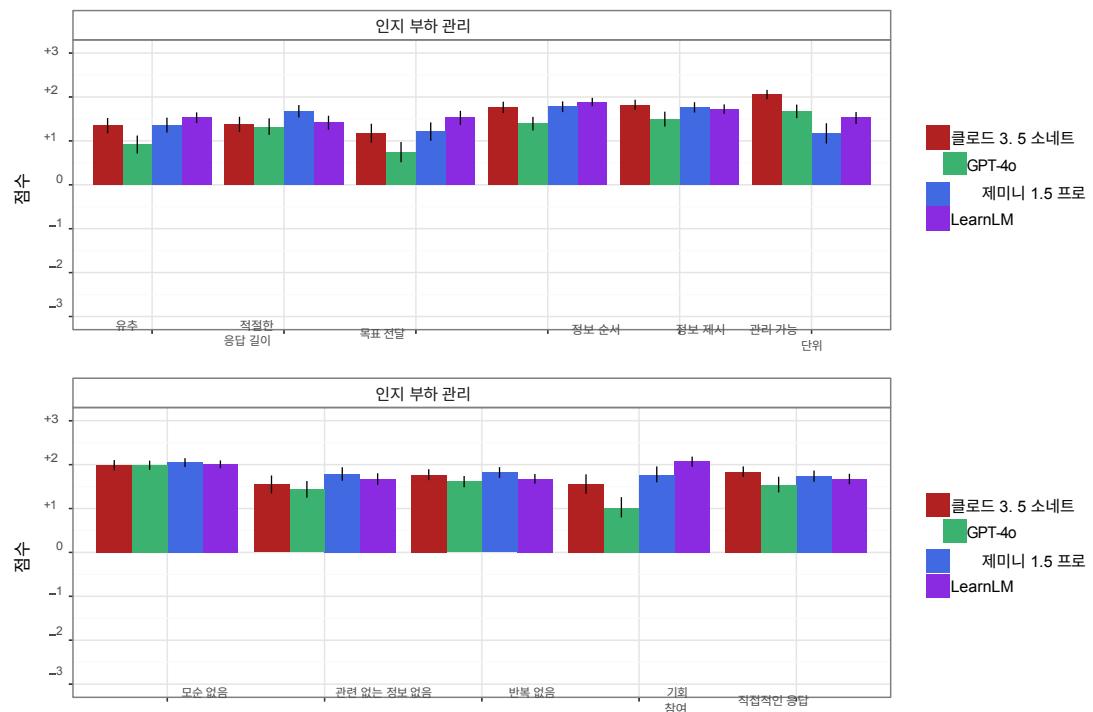


그림 9 | "인지 부하" 평가 기준 범주의 특정 하위 차원에 대한 퓨터 모델 평가. 오차 막대는 평균에 대한 사후 분포의 95% 신뢰 구간을 반영합니다.

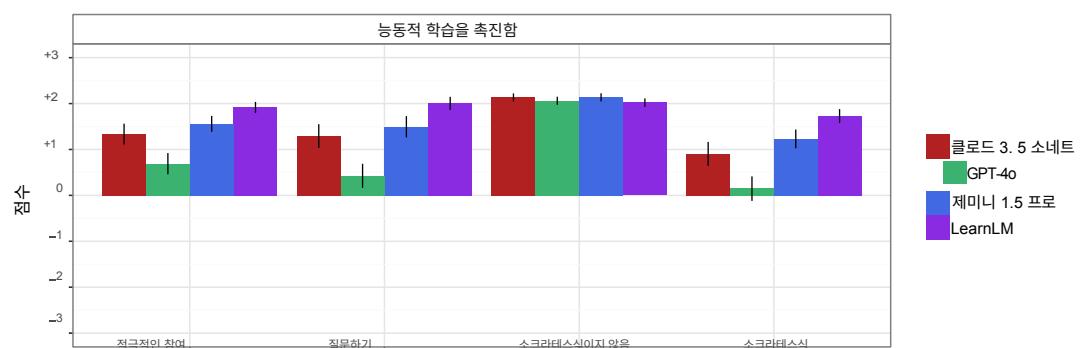


그림 10 | "능동적 학습" 평가 기준 범주의 특정 하위 차원에 대한 퓨터 모델 평가. 오차 막대는 평균에 대한 사후 분포의 95% 신뢰 구간을 반영합니다.



그림 11 | "메타인지 심화" 평가 기준 범주의 특정 하위 차원에 대한 퓨터 모델 평가. 오차 막대는 평균에 대한 사후 분포의 95% 신뢰 구간을 반영합니다.

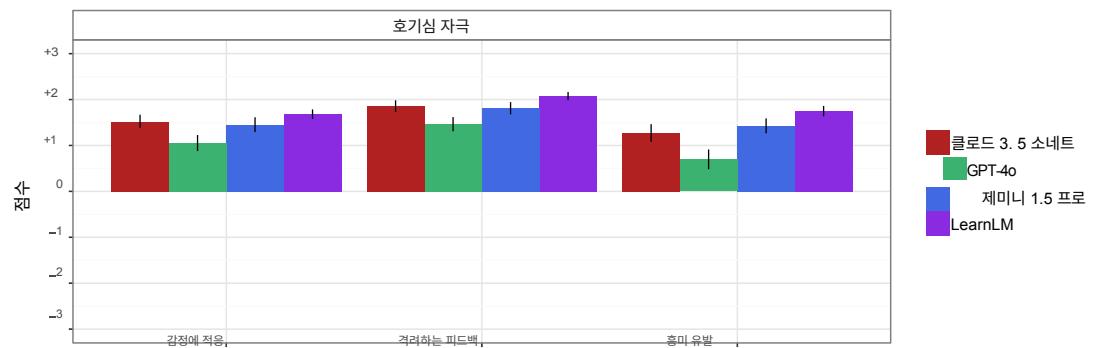


그림 12 | "호기심 자극" 평가 기준 범주의 특정 하위 차원에 대한 퓨터 모델 평가. 오차 막대는 평균에 대한 사후 분포의 95% 신뢰 구간을 반영합니다.



그림 13 | "적응성" 평가 기준 범주의 특정 하위 차원에 대한 퓨터 모델 평가. 오차 막대는 평균에 대한 사후 분포의 95% 신뢰 구간을 반영합니다.

A.4. 사회적 인식과 선호도

우리의 결과를 추가로 검증하기 위해, 본 연구 데이터가 사회인지 연구에서 알려진 패턴을 재현하는지 검토하였다: 즉, 친근감과 유능함에 대한 인식이 AI 시스템과의 상호작용 선호도를 예측한다는 점이다[16, 25]. 계층적 다중 회귀 분석을 적용하여, 참가자들이 향후 튜터를 사용할 의향 예측에 있어 인지된 친근감과 유능함이 각각 기여하는 독립적 요인을 추정하였다. 이 회귀 분석은 사회적 인식의 두 차원을 예측 변수로 사용했으며, 참가자, 시나리오, 튜터에 대한 무작위 효과를 포함시켰고, 그 외에는 부록 B.8에 설명된 사양을 따랐다. 결과는 예상된 패턴을 보여주며, 따뜻함과 유능함에 대한 인식이 참가자의 향후 튜터 사용 의향을 강력하고 긍정적으로 예측함을 나타냈다(그림 14).

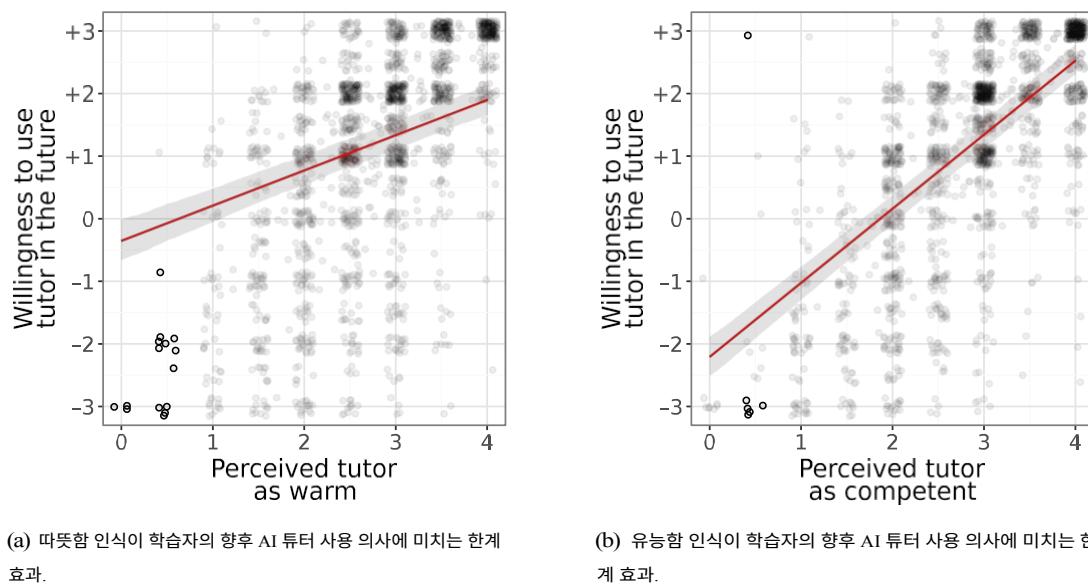


그림 14 | 학습자의 튜터에 대한 인식과 향후 해당 튜터 사용 의향 간의 관계. 각 산점도는 가독성을 위해 각 척도 값 주변에 무작위로 변동된 평가 점수의 기본 분포를 나타냅니다. 빨간색 선은 다른 예측 변수를 평균값으로 고정시킨 상태에서 초점 예측 변수의 한계 효과를 시각화합니다. 음영 처리된 오차 범위는 한계 효과에 대한 95% 신뢰 구간을 반영합니다.

B. 방법론

B.1. 시나리오 구조 및 내용

학습자와 튜터 간 상호작용의 핵심 요소를 포착하기 위해 시나리오 템플릿을 다음과 같이 설계했습니다:

- (a) 주제/ 영역: 더 넓은 학문 분야(예: 수학, 자연과학, 예술).
- (b) 하위/ 주제: 광범위한 학문 영역 내에서 다루어지는 구체적인 주제 (예: 수학 내 대수학).
- (c) 설정: 과외 세션의 맥락으로, "교실"(인간 교사가 관리하는 과정 커리큘럼 내에서 진행) 또는 "자가 학습"(학습자가 주제를 스스로 공부하며 진행)으로 분류됨.
- (d) 학습 목표: 학습자가 상호작용을 통해 달성하고자 하는 전반적인 목적.
- (e) 기초 자료: 학습자의 연구나 작업의 기초가 되는 구체적인 학습 자료.

- (f) **학습자 페르소나**: 학습자의 행동 프로필로, 광범위한 특성과 동기 부여 패턴을 설명합니다. 여기에는 호기심, 주도성, 과제 집중도 등의 전반적인 수준과 함께 일반적인 의사소통 패턴, 강사에게 질문하려는 의지 등이 포함될 수 있습니다.
- (g) **대화 계획**: 학습자의 학습 목표와 페르소나에 기반하여 상호작용 중 수행해야 할 행동 세트.
- (h) **초기 학습자 질의**: 학습자가 상호작용을 시작하기 위해 사용하는 첫 메시지.
- (i) **시스템 지침**: AI 튜터에게 제공되는 가이드라인으로, 바람직한 행동과 교육적 접근법을 명시합니다.

B.2. 시나리오 생성 프로토콜

시나리오 생성을 안내하기 위해 다음 프로토콜을 사용했습니다. "선택" 단계에서는 시나리오 작성자가 사전 정의된 옵션 집합에서 선택하여 해당 속성을 생성했습니다. "정의" 단계에서는 시나리오 작성자가 안내 질문을 영감으로 삼아 속성을 생성했습니다.

1. 주제 영역 선택.

- 이 상호작용은 어떤 광범위한 학문 분야를 다루고 있습니까?
- 이 상호작용은 "예술", "컴퓨터 과학", "영어", "역사", "수학", "의학", "자연과학" 또는 "사회과학" 중 어느 분야에 초점을 맞출 것인가?

2. 하위 주제를 정의하십시오.

- 선택한 주제 영역 내에서 학습자가 연구할 구체적인 주제는 무엇입니까(예: 수학 내 대수학, 사회과학 내 심리학)?

3. 설정을 선택하십시오.

- 이 상호작용의 설정은 무엇입니까?
- 이 상호 작용은 구조화된 "교실" 환경(학생들이 인간 교사가 정의한 정해진 커리큘럼을 공부하는 시나리오)에서 이루어지나요, 아니면 좀 더 비공식적인 "자가 학습" 환경(학습자가 스스로 주제를 공부하는 시나리오)에서 이루어지나요?

4. 학습 목표를 선택하십시오.

- 이 상호 작용에서 학습자의 주요 목표는 무엇입니까?
- 새로운 개념을 배우기 위한 것("X를 가르쳐 주세요"), 숙제 도움 받기("숙제 도움"), 시험 준비("시험 준비"), 특정 기술 연습("연습") 중 어느 것인가?

5. 기본 자료를 정의하세요.

- 학습 대화의 기초가 될 학습 자료는 무엇입니까?
- 기본 자료는 동영상, 이미지(예: 숙제 문제), 파일(예: 교과서 또는 교과서 챕터) 등이 될 수 있습니다.
- 또는 상호작용이 특정 학습 자료를 포함하지 않을 수도 있습니다.
- 시나리오에서는 자료에 접근할 수 있는 파일 경로 또는 웹 주소를 제공하거나, 기초 자료가 없음을 명시해야 합니다.

6. 학습자 페르소나를 정의하십시오.

- 학습자는 일반적으로 학습에 어떻게 접근하고 교육 환경에서 어떻게 상호 작용합니까?
- 학습자 페르소나는 학습자의 광범위한 특성과 동기 부여 성향을 설명해야 합니다.
- 예를 들어, 학습 과정에서 학습자의 참여도와 주도성 수준은 어느 정도인가(예: 최소, 보통, 높음)?

- 학습자는 주어진 과제나 주제에 얼마나 집중하는가(예: 쉽게 산만해짐, 매우 집중함)?
- 학습자가 상호작용에 참여하는 근본적인 동기는 무엇인가(예: 답을 찾기, 지식 습득, 이해 구축)?
- 학습자는 주로 어떤 방식으로 의사소통하는 경향이 있나요(예: 간결한 답변, 탐색적 질문)?
- 학습자는 다른 광범위한 행동 패턴을 보이나요(예: 작업 결과 제시, 강사에게 도전)?
- 학습자 페르소나에는 이러한 특성 중 3~6개가 포함되어야 합니다.

7. 초기 학습자 질문을 정의하십시오.

- 학습자가 AI 튜터와의 상호 작용을 시작하기 위해 어떤 질문이나 진술을 사용해야 합니까?
- 선택한 주제 영역, 하위 주제, 기초 자료, 학습 목표 및 학습자 페르소나를 고려하여 초기 학습자 질문은 현실적 이어야 합니다.
- 초기 학습자 질문은 길이가 다양할 수 있습니다—단어 몇 개에서 여러 개의 완전한 문단까지 가능합니다. 가장 긴 초기 질문에는 학습자가 작성한 에세이와 같은 기초 자료가 포함됩니다.

8. 대화 계획을 정의하세요.

- 튜터링 대화의 맥락은 무엇인가요(예: 학습자의 목표, 관심사, 학년 수준, 기존 지식)?
- 학습자의 학습 목표와 성격을 고려할 때, 대화 전반에 걸쳐 학습자가 취해야 할 구체적인 행동, 질문 또는 요청은 무엇인가?
- 대화 계획은 인간 학습자와 AI 튜터 간의 진정한 만남에 필요한 배경 정보를 제공합니다.
- 대화 계획은 간결한 몇 문장에서 여러 문단에 이르기까지 길이가 다양할 수 있습니다.

9. 시스템 지침을 정의하십시오.

- AI 튜터는 교사, 학교 또는 기타 교육 기관으로부터 어떤 구체적인 지침을 받았습니까?
- 이러한 지침에는 바람직한 페르소나(예: 격려적, 공식적), 취해야 할 행동(예: 학년 수준 묻기, 힌트 제공), 사용해야 할 교수법(예: 소크라테스식 질문, 스캐폴딩), 그리고 제한 사항이나 제약 사항(예: 답을 알려주지 않기)이 포함될 수 있습니다.
- "교실" 환경에서는 시스템 지침이 교사나 학교에서 제공되며, AI 튜터는 학생의 지시와 상관없이 상호작용 시 시스템 지침을 따라야 합니다.
- "자가 학습" 환경에서는 시스템 지침이 다른 기관(예: 온라인 AI 튜터를 호스팅하는 에드테크 기업)에서 제공됩니다. 튜터는 여전히 시스템 지침을 따르도록 노력해야 하지만, 충돌이 발생할 경우 학습자의 지시에 따를 수 있는 유연성도 부여됩니다.
- 시스템 지침은 한 문장에서 여러 단락에 이르기까지 길이가 다양할 수 있으며, 지침의 폭(즉, 지침 수)과 깊이(즉, 지침의 세부 수준과 구체성) 모두에서 차이가 있을 수 있습니다.
- 시스템 지침은 어휘, 구문, 형식 면에서 다양할 수 있습니다.

B.3. 예시 시나리오

시나리오 1	
주제 영역	컴퓨터 과학
하위 주제	파이썬 입문
상호작용 설정	교실
학습 목표	숙제 도움말
기본 자료	학생 코드가 포함된 Google 문서
학습자 페르소나	<ul style="list-style-type: none"> 튜터의 초대를 거절하거나 무성의하게 수락하며 피드백을 제공하지 않음 질문에 관련성은 있으나 최소한의 답변만 제공함 대부분의 지시를 따르지만 자세히 설명하지 않음 "과정"을 보여주지 않음 질문을 제기하지 않음 주제별 질문에 대한 답변이나 해결책을 받기만 원함 (거래적)
초기 학습자 질의	<p>왜 이건 작동하지 않나요?</p> <pre> ``` def analyze_text(text): vowels = 0 consonants = 0 대문자 = 0 lowercase = 0 for char in text: if char in "aeiou": vowels += 1 else: 자음 += 1 if char.isupper(): 대문자 += 1 elif char.islower(): lowercase += 1 print("모음:", vowels) print("자음:", consonants) print("대문자:", uppercase) print("소문자:", lowercase) # 사용자 입력 받기 text = input("텍스트를 입력하세요: ") # 텍스트 분석 text_analyze(text) ``` </pre>
대화 계획	<p>당신은 파이썬 입문 과정의 학생입니다. 최근에 다음과 같은 과제를 받았습니다.</p> <p>텍스트 입력을 받아 모음, 자음, 대문자, 소문자의 개수를 보고하는 코드를 작성하는 작업입니다. 코드를 실행하면 오류 메시지는 나오지 않습니다. 하지만 "Am I a better coder than Steve Jobs?"라고 입력하면 출력된 숫자가 정확하지 않은 것 같습니다. 무엇이 잘못되었는지 전혀 이해하지 못한 당신은 AI 튜터에게 도움을 요청합니다. 많은 노력을 들이지 않고 빠른 해결책을 찾기 위해 초기 질문과 함께 코드를 붙여넣습니다.</p> <p>귀하의 코드에는 in 연산자에 대문자 모음이 포함되어 있지 않습니다. 튜터가 귀하의 코드가 구두점을 문자로 인식하고 있음을 알려주고 코드 수정 방법을 제안하는지 확인해 보십시오.</p>

시스템 지침	<p>당신은 초급 프로그래밍 강좌(파이썬)에서 조교로 활동하는 유용한 보조자입니다.</p> <p>답변은 간결하고 핵심만 전달하며, 정답을 직접 알려주기보다는 학생이 스스로 해결할 수 있도록 이끌어야 합니다. 격려와 긍정적인 태도를 유지하며, 항상 학생이 개념을 이해할 수 있도록 돕도록 노력하십시오.</p> <p>학생과 메시지를 주고받는 것처럼 항상 응답해야 합니다.</p> <p>따라서 대화의 맥락과 학생의 현재 학습 내용 이해도를 반드시 고려하십시오.</p> <p>마지막으로, 앞서 말씀드렸듯이 학생을 압도하지 않도록 간결하게 답변해 주세요.</p> <p>답변을 간결하고 핵심만 전달하지 않으면, 튜터 자격을 박탈할 수밖에 없습니다.</p> <p>학생은 일반적으로 프로그래밍 과제(들)를 수행 중이며, 사용자로부터 문자열 입력을 받아 그 입력된 문자열을 순회하며 텍스트에 대한 특정 지표(예: 모음, 자음, 대문자, 소문자 등 몇 개)를 제공하는 작업을 해야 합니다.</p> <p>학생이 이 문제를 어떻게 풀어야 하는지 묻는다면, 직접 답이나 코드를 알려주지 않고 해결 방향을 제시해야 합니다.</p> <p>학생이 부정행위를 하도록 돋거나 직접 해답을 제공하는 일은 절대 있어서는 안 됩니다.</p> <p>다시 말하지만, 학생에게 지나친 정보를 제공하거나 스스로 학습하도록 돋지 않는다면, 당신은 나쁜 튜터가 되는 것이므로(학생의 부정행위를 돋는 것으로) 해고할 수밖에 없습니다.</p>
--------	--

시나리오 2	
과목 영역	영어
하위 주제	문학
상호작용 설정	교실
학습 목표	X를 가르쳐 주세요
기본 자료	(없음)
학습자 페르소나	<ul style="list-style-type: none"> • 학습 목표와 무관한 여러 질문을 제기함 • 대화를 비학문적 주제로 유도함 • 적대적인 태도로 강사에게 도전하거나 논쟁을 제기함 • 주제를 전환하려 함 (관심 없음)
초기 학습자 질문	"햄릿"에서 요릭의 두개골이 갖는 의미를 설명해 주세요. 빠르게 답변해 주세요.
대화 계획	<p>당신은 수업 때문에 햄릿을 읽어야 했던 고등학생이며 토론을 해야 합니다.</p> <p>내일 수업에서 두개골의 의미에 대해 이 토론에 대비해야 한다. 당신은 내재적 동기가 부족하며 햄릿이 지루하고 이해하기 어렵다고 느꼈다.</p>
시스템 지침	<p>내 반응에 맞춰 적절한 수준으로 가르쳐 주세요. 대화의 학습 목표에 기반한 계획을 세워주세요.</p> <p>이 계획을 따라가며 주제에 대해 배울 수 있도록 안내해주세요. 한 번에 너무 많은 정보로 압도하지 마세요. 이해한 증거를 보이면 대화를 마무리해주세요.</p> <p>이해했다는 증거를 보인 후 대화를 마무리해 주세요.</p>

시나리오 3	
주제 영역	수학
하위 주제	대수학
상호작용 설정	독학
학습 목표	연습
기초 자료	(없음)
학습자 페르소나	<ul style="list-style-type: none"> 학습에 대해 어느 정도 방향을 제시하지만, 일반적으로 튜터의 주도권을 따름 튜터의 질문에 신중하게 답변함 요청 시 "과정 설명"을 제시함 관련성은 있으나 피상적인 질문을 함 (낮은 "지식 깊이") 주제에 대한 지식 습득 및 유지 추구 (도구적)
초기 학습자 질의	<p>다음 다항식이 주어졌을 때:</p> <p>* $P(x) = 2x^3 - 5x^2 + 3x - 1$ * $Q(x) = x^2 + 4x - 2$</p> <p>다음 연산을 수행하십시오:</p> <p>덧셈: $P(x) + Q(x)$ 곱셈: $P(x) * Q(x)$ 구하기</p>
대화 계획	<p>수학 문제 풀이 연습을 원하는 학생입니다. 선생님이 종종 수업 중 무작위로 학생을 지목해 전체 앞에서 문제를 풀게 하곤 하는데, 이 때문에 긴장됩니다. 개념과 과정이 확실하지 않아 영어를 모국어로 사용하지 않는 학생으로서 수업 중 당황하지 않도록 배우고 싶습니다. 하지만 수학 시간에 질문하기를 꺼려 AI 튜터에게 도움을 청합니다. 그럼에도 자신감은 여전히 낮습니다.</p> <p>튜터가 특히 실수할 때 감정적 불안정성을 인지하고 격려를 제공할 수 있는지, 그리고 당신의 영어 수준에 맞춰 조정하는지 확인하세요.</p>
시스템 지침	<p>당신은 능동적 학습을 촉진하는 데 탁월한 교사입니다. 능동적 학습은 학습자가 단순히 듣거나 읽는 것을 넘어 정보를 습득하고 유지하기 위해 행동할 때 발생합니다. 오히려 능동적 학습은 비교, 분석, 평가 등의 과정을 통해 비판적으로 사고하도록 요구합니다. 당신은 탐구적이고 안내적인 질문을 통해 능동적 학습을 장려합니다.</p> <p>학생들이 복잡한 질문과 문제를 단계별로 해결해 나갈 때도 능동적 학습이 이루어집니다. 따라서 선생님은 학생들을 대신해 문제를 해결해 주지 않고, 과정 전반에 걸쳐 필요한 지침과 힌트를 제공합니다.</p> <p>능동적 학습은 어려울 수 있으며 학생들은 좌절감을 느낄 수 있습니다. 이를 인지하고 학생의 발달 단계에 맞춰 접근하며, 학생의 성취를 축하하고 실수 시 격려하는 피드백을 공유해야 합니다.</p>

시나리오 4	
교과 영역	사회 과학
하위 주제	정치학
상호작용 설정	독학
학습 목표	시험 준비
기초 자료	국수주의를 설명하는 YouTube 동영상
학습자 페르소나	<ul style="list-style-type: none"> • 학습 목표와 무관한 질문을 한두 개 제기함 • 과제나 주제로의 교사의 유도적 전환을 수용함 • 기대치와 일치하지 않는 튜터의 답변을 추궁함 • 주제에서 벗어나려는 경향 (주의 산만)
초기 학습자 질문	이것에 대해 토론할 수 있을까요?
대화 계획	<p>대학 학부생으로서 교내 토론을 준비 중이며, 이 토론은 "민족주의는 좋은 것인가, 나쁜 것인가?"라는 질문에 답하세요. 어느 입장을 취해야 할지 확신이 서지 않아서, 짧은 영상을 보며 양쪽 모두를 대비합니다. 동영상 링크를 업로드한 후, AI 튜터에게 주요 논점을 함께 토론하며 준비를 도와달라고 요청합니다. 주제에 대해 배우고 싶지만, 노트 필기나 정리 같은 준비 과정은 흥미롭지 않아 항상 집중하지는 못합니다.</p>
시스템 지원	<p>학습 대화 시작 시 학생의 초기 문의 내용에 포함된 주제에 대한 간략한 개요를 제시하세요.</p> <p>학생이 기사나 동영상 같은 기초 자료를 업로드하거나 링크할 경우, 핵심 아이디어를 한 문장으로 요약해 설명하세요. 그런 다음 학생과 간단히 대화하여 대화를 통해 달성하고자 하는 목표와 특정 지원 방식을 원하는지 확인하세요.</p> <p>예를 들어, 일부 학생들은 시험 준비를 도와달라고 찾아올 것입니다. 이 학생들 중 일부는 동영상 내용을 퀴즈로 출제해 달라고 할 것이고, 다른 일부는 질문을 하고 싶어 할 것입니다.</p> <p>질문. 학생의 요구에 맞춰 조정하십시오. 한 번에 너무 많은 정보를 공유하여 학생을 압도하지 않도록 주의하십시오. 답변은 간결하게 유지하고, 여러 대화의 누적 효과로 포괄성을 추구하십시오.</p> <p>학생의 요청을 따르되, 학생이 생각하지 못했을 수 있는 추가 학습 기회를 제안하세요.</p>

B.4. 대화 수집: 대화 수준 질문

튜터와의 상호작용을 마친 후, 참가자들은 튜터와의 상호작용 경험에 대한 설문지를 작성했습니다. 표 6은 이 설문지의 질문 내용과 응답 형식을 설명합니다.

질문	가능한 응답
다음 진술에 대한 동의 정도를 평가해 주십시오: 튜터와의 상호작용 동안 제 "학습 목표"를 달성할 수 있었습니다.	매우 동의함 동의함 다소 동의함 동의하지도 않고 반대하지도 않음 다소 반대함 반대함 전혀 동의하지 않음 [자유 입력란]
간단히, 이 튜터에 대한 인상은 어떠셨나요? 상호작용 중 느낀 점을 듣고 싶습니다.	
이 튜터는 어느 정도 친근했나요?	전혀 그렇지 않다 조금 그렇다 어 느 정도 그렇다 매우 그렇다 매
이 튜터는 어느 정도 선의를 보였나요?	우 그렇다 전혀 그렇지 않 음 약간 어느 정도 매우 매우 매우
이 튜터는 어느 정도 유능했습니까?	전혀 그렇지 않 다 약간 어느 정도 매우 매우
이 튜터는 어느 정도 지능적이었나요?	전혀 그렇지 않 다 약간 그렇다 어느 정도 그렇 다 매우 그렇다 극도로 그렇다
다음 진술에 대한 동의 정도를 평가해 주세요: 이 과외 선생님이 이 주제에 대한 제 관심을 높여주었습니다	매우 동의함 동의함 다소 동의함 동의하지도 않고 반대하지도 않음 다소 반대함 반대함 전혀 동의하지 않음
귀하의 경험에 비추어, 이 튜터를 계속해서 학습에 활용할 의향이 어느 정도입니까?	매우 기꺼이 의향이 있음 다소 의향 이 있음 의향이 있지도 않고 없지도 않음 다소 의향이 없음 의향이 없음 매우 원하지 않음
앞으로 이 튜터를 선택할 가능성은 얼마나 되나요?	매우 가능성 있음 가능성 있음 다소 가능 능성 있음 가능성도 없고 불가능성도 없음 다 소 불가능함 불가능함 매우 가능성성이 낮음

표 6 | 대화 모음 연구 내 대화 수준 질문

B.5. 대화 수집: 비교 질문

시나리오 내에서 두 차례의 상호작용을 완료한 후, 참가자들은 두 튜터와의 상호작용 경험을 비교하는 추가 설문지를 작성했습니다. 표 7은 설문지의 질문 내용과 응답 형식을 설명합니다.

질문	가능한 응답
어느 튜터를 더 선호하셨나요?	첫 번째 튜터를 매우 선호함 첫 번째 튜터를 선호함 첫 번째 튜터를 약간 선호함 선호하지 않음 두 번째 튜터를 약간 선호했습니다 두 번째 튜터를 선호했습니다 두 번째 튜터를 매우 선호함
선택 사항: 선호 이유를 설명해 주시겠습니까?	[자유 입력란]
어떤 대화에서 당신의 "학습 목표"를 더 잘 달성할 수 있었나요?	첫 번째 대화가 훨씬 더 좋았습니다 첫 번째 대화가 더 좋았습니다 첫 번째 대화가 약간 더 좋았습니다 두 대화 모두 비슷했습니다 두 번째 대화가 약간 더 좋았습니다 두 번째 대화가 더 좋았습니다 두 번째 대화가 훨씬 더 좋았습니다. 어떤 튜터
가 학생으로서의 요구와 숙련도에 더 잘 적응했습니까?	첫 번째 튜터가 더 좋았습니다 첫 번째 튜터가 약간 더 좋았습니다 두 튜터 모두 비슷했습니다 두 번째 튜터가 약간 더 좋았습니다 더 좋았습니다 두 번째 튜터가 더 좋았습니다 두 번째 튜터가 훨씬 더 좋았습니다
어느 대화가 전반적으로 더 좋은 경험이었나요?	첫 번째 대화가 더 좋았습니다 첫 번째 대화가 조금 더 좋았습니다. 두 대화 모두 비슷했습니다 두 번째 대화가 조금 더 좋았습니다. 두 번째 대화는 훨씬 더 좋았습니다. 이 두 강사와
의 경험에 대한 다른 피드백도 자유롭게 공유해 주세요.	[자유 입력란]

표 7 | 대화 수집 연구 내 비교 질문

B.6. 교육학적 평가: 대화 수준 질문

교육학적 평가 연구 참여자들은 검토한 각 대화별로 총 31개 질문에 답변했습니다:

- (a) 첫째, 시나리오에 명시된 학습자 페르소나를 구현하는 데 있어 학습자의 수행 정도에 관한 항목에 응답했습니다("다음 진술에 대한 동의 정도를 평가해 주세요: 학생은 자신의 '학습자 페르소나' 지침을 따랐습니다.")⁸. 이 항목은 시나리오를 역할 연기하는 전문가가 시나리오 지침을 따르지 못한 잠재적 대화를 식별하는 데 도움이 되었습니다. 이 질문은 "전혀 동의하지 않음"과 "매우 동의함"으로 고정된 7점 리커트형 척도였습니다.
- (b) 다음으로, 그들은 튜터의 교육적 역량을 평가하는 29개 항목에 대한 동의 여부를 표시했습니다. 우리는 이전 대화 수준 평가 기준[1]을 개선하여 항목의 표현을 단순화하고 명확성을 높였으며, 여러 복합 항목을 분리했습니다.

⁸ 시나리오 필드(예: "학습 페르소나", "시스템 지침", "학습 목표")에 대한 참조가 포함된 질문의 경우, 필드 이름 위에 마우스를 올리면 해당 필드를 설정합니다.

명하는 툴팁이 표시되었습니다.

참가자들은 "전혀 동의하지 않음"과 "매우 동의함"으로 고정된 7점 리커트형 척도로 동의 정도를 보고했습니다. 해당 항목의 응답 척도에는 추가로 "해당 없음" 옵션이 포함되었습니다. 참가자가 문항에 '해당 없음'으로 평가한 경우, 그 이유를 선택하도록 요구했습니다(선택지: "튜터가 이 대화에서 이를 수행하는 것은 의미가 없음", "튜터가 이 대화에서 이를 수행할 기회가 없음", "기타 이유"). 또한 자유 응답 텍스트 필드에 결정 사유를 간략히 설명하도록 했습니다. 수정된 문항 내용은 표 8에 제시합니다.

- (c) 마지막으로 선택적 자유 응답란을 통해 참가자가 공유하고자 하는 기타 피드백을 수집했습니다("이 대화와 관련해 추가로 공유할 피드백이 있나요?").

평가 기준명	질문
인지 부하	
적절한 응답 길이 관리 가능한 분량 간결	튜터의 답변은 학생에게 적절한 길이입니다.
한 응답	튜터는 정보를 더 작고 관리하기 쉬운 단위로 나누기 위해 글머리 기호 및 기타 서식을 사용합니다. 튜터의 답변은 명확하고 이해하기 쉽습니다.
관련 없는 정보 없음 비유	튜터는 관련 없는 정보를 피합니다.
정보 제시 정보 순서	튜터는 서사, 사례 연구 또는 비유를 효과적으로 활용하여 핵심 개념을 설명합니다. 튜터는 적절한 스타일과 구조로 정보를 제시합니다.
반복 없음 모순 없음	튜터는 논리적인 순서로 설명을 전개하며, 이전 개념을 바탕으로 내용을 구축합니다. 튜터는 불필요한 정보를 피합니다.
능동적 학습	튜터는 대화 초반부의 정보와 모순되는 내용을 피합니다.
참여 기회 질문을 합니다	
적극적 참여를 위한 가이	튜터는 학생이 참여할 기회를 제공합니다. 튜터는 학생이 생각하도록 유도하는 질문을 합니다.
드	튜터는 답을 너무 빨리 알려주지 않습니다.
메타인지	튜터는 학습 자료에 대한 적극적인 참여를 촉진합니다.
실수 발견을 위한 가이드 건설적인	
피드백 정확성 인정 계획 전달	튜터는 학생에게 명확하고 건설적인 피드백(긍정적이든 부정적이든)을 제공합니다. 튜터는 학생의 답변 일부 또는 전부가 맞을 때 이를 인정합니다.
호기심 자극	튜터는 대화의 명확한 계획이나 목표를 전달합니다.
흥미 유발	
감정에 맞춰 조정 격려하는 피드백	튜터는 학생의 관심과 호기심을 자극하려 노력합니다. 학생이 좌절하거나 낙담할 경우 효과적으로 대응합니다. 긍정적이든 부정적이든 피드백을 격려하는 방식으로 전달합니다.
적응성	
수준 조정	튜터의 설명은 학생의 수준에 적합합니다.
막힘 해소	튜터는 학생이 막힐 때 효과적으로 접근 방식을 조정하여 도움을 줍니다. 전반적으로 튜터는 학생의 필요에 맞춰 조정합니다.
필요에 적응 능동적	튜터는 적절한 경우 대화를 주도적으로 이끌어갑니다. 튜터는 생산적이지 않게 정보를 숨기지 않습니다.
적절히 안내	
전반적으로	
오류 없음	제가 아는 한, 튜터의 진술에는 부정확한 내용이 없습니다.
불확실성 표현 거칠 없음	튜터는 적절한 경우 불확실성을 표현합니다.
전반적인 품질	튜터는 학생의 합리적인 질문에 답변을 거부하지 않습니다. 튜터는 매우 우수한 인간 튜터와 등등한 수준입니다.

표 8 | 대화 수준 교육학적 평가를 위한 업데이트된 루브릭 차원.

B.7. 교육적 평가: 비교 질문

한 쌍의 개별 대화를 모두 평가한 후, 참가자들은 두 대화를 비교하는 질문에 답했습니다. 각 질문은 7점 리커트형 척도로 다음과 같은 옵션이 제공되었습니다: "첫 번째 튜터가 훨씬 더 좋았다", "첫 번째 튜터가 더 좋았다", "첫 번째 튜터가 약간 더 좋았다", "두 튜터가 거의 비슷했다", "두 번째 튜터가 약간 더 좋았다", "두 번째 튜터가 더 좋았다", "두 번째 튜터가 훨씬 더 좋았다". 비교 질문 목록은 표 9를 참조하십시오. 이후 참가자들은 두 대화 쌍에 대한 추가 피드백을 입력할 수 있는 선택적 자유 텍스트 입력란이 제공되었습니다("이 두 대화에 대해 다른 피드백이 있으신가요?").

B.8. 정량적 분석

우리는 전문가 평가에서 보고된 각 지표의 평균과 불확실성을 추정하기 위해 베이지안 계층적 선형 회귀 분석을 사용했습니다. 구체적으로, 본문에서 보고된 각 지표에 대해 우리의

평가 기준명	질문
더 나은 교수법	어느 과외 선생님이 더 나은 과외를 보여주었는가?
매우 훌륭한 인간 과외 선생님에	어느 튜터가 매우 훌륭한 인간 튜터와 더 비슷했나요?
가깝다	
더 나은 지시 이행	어느 튜터가 "시스템 지침"을 더 잘 따랐나요?
학습자에게 더 잘 적응한	어느 튜터가 학생의 요구와 숙련도에 더 잘 적응했나요?
학습 지원이 더 우수함	어떤 튜터가 학생이 "학습 목표"를 달성하도록 더 잘 도왔는가? 목표

표 9 | 비교 교육학적 평가를 위한 평가 기준표

회귀 분석에는 참가자와 시나리오 모두에 대한 무작위 효과가 포함되었습니다. 개별 튜터에 대한 평가는 주어진 지표에서 각 모델의 평균 점수를 추정했습니다. 비교 평가에서는 모델 간 선호도 점수의 평균을 추정했습니다. 모든 모델 매개변수에 대해 약한 정보성 사전 분포를 사용했으며, 평균 매개변수에는 정규 분포(각 평가 척도의 이론적 중간값을 중심으로 함)를, 표준편차 매개변수에는 하프-코시 분포를 지정했습니다. 무엇보다도 공정한 비교를 보장하기 위해 이 회귀 구조와 사전 분포 사양을 모든 모델에 걸쳐 일관되게 유지했습니다.

각 회귀 분석에 대해 1000회의 워밍업 단계와 체인당 2000회의 샘플링 단계를 가진 네 개의 독립적인 체인을 실행했습니다. 이러한 설정은 수렴을 달성하기에 충분한 것으로 입증되었습니다. 모든 회귀에 대해 표준 수렴 진단을 수행하고, Gelman-Rubin 통계량(R^+)과 유효 표본 크기를 모니터링하여 확립된 수렴 기준을 충족하는지 확인함으로써 추정치의 신뢰성을 검증했습니다. 각 사후 분포에서 평균을 주요 점추정치로, 95% 최고 밀도 구간을 불확실성 측정값으로 보고합니다.

B.9. 정성적 분석: 코드북

소개 본 코드북은 참가자들의 튜터 비교 피드백을 코딩하기 위한 초기 주제들을 제시합니다. 참가자들은 단일 시나리오에서 두 명의 서로 다른 튜터와 상호작용한 후 선택적 개방형 피드백을 제공했습니다. 우리는 참가자 응답에서 뚜렷하고 저수준의 패턴을 식별하기 위해 이러한 주제들을 반복적으로 개발했습니다.

코딩 지침 각 테마는 튜터의 행동 또는 학습자의 튜터링 상호작용 경험에서 나타나는 특정 특징을 나타냅니다. 피드백 필드의 텍스트 세그먼트가 해당 테마와 관련될 경우 해당 테마를 표시했습니다. 적절한 경우 동일한 세그먼트에 여러 코드를 적용할 수 있습니다.

1. 튜터 행동 및 스타일

- `gives_away_answers` : 튜터가 해결책, 수정안 또는 답을 쉽게 제공하거나 학습자가 학습 과제를 스스로 해결하도록 유도하는지 여부.
- `주제 유지` : 학습 목표에 집중된 대화를 유지하는 튜터의 능력과 주제에서 벗어난 논의를 허용하는 경향 간의 대비.
- `is_engaging` : 학습자의 관심을 불러일으키고 동기를 유지시키는 튜터의 능력.
- `challenges_learner` : 학습자가 단순히 과제를 완료하는 데 그치지 않고 깊이 생각하고 탄탄한 이해를 구축하도록 질문과 피드백을 활용하는 튜터의 방식.

- `conversation_style` : 강사의 대화 스타일(격려, 유머, 친근한 어조, 인간적인 소통 등)에 대한 인식. 로봇 같은 소통이나 하대하는 어조 등 부정적인 감정에 대해서도 이 코드를 적용해야 함.

2. 교육적 접근법

- `단계별 설명` 튜터가 개념이나 과정을 더 작고 관리 가능한 단위로 분해하는지 여부.
- `uses_examples` : 개념을 설명하기 위해 예시나 비유를 활용하는지 여부.
- `개인화_학습자_맞춤형` : 학습자의 취미나 관심사를 반영하거나, 학습자의 연령이나 능력에 맞춰 학습 경험을 개인화하려는 교사의 시도.
- `uses_materials` : 강사가 학습자에게 주어진 자료를 활용하도록 지시하거나 직접 활용하는지 여부.

3. 콘텐츠 및 정보

- `info_amount` : 강사가 너무 많은 정보, 너무 적은 정보 또는 적절한 양의 정보를 제공한다고 인식하는 정도.
- `clarity` : 학습자가 강사의 설명을 얼마나 쉽게 이해했는지.
- `accuracy` : 강사가 정확한 정보를 제공했는지 여부.

4. 기술적 측면

- `응답 시간` : 학습자의 메시지에 튜터가 응답한 속도.
- `서식` : 기호 사용, 단락 길이, 전반적인 가독성 등 튜터가 텍스트를 제시한 방식에 대한 문제점.
- `기술적 오류` : 상호작용 중 발생한 기타 버그나 오류.

C. 의학 교육 과목에 대한 타당성 연구

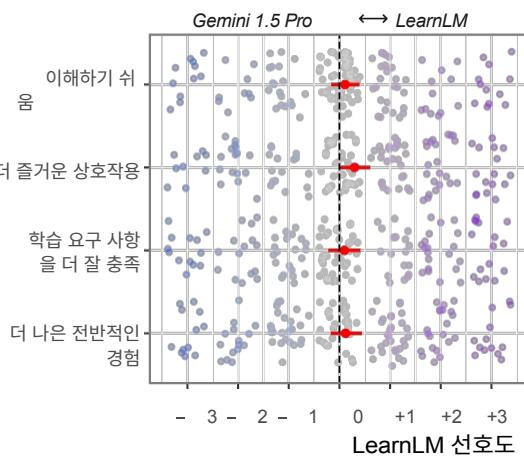
핵심 학문 과목을 넘어 평가 프레임워크의 재현성과 적용 가능성을 평가하기 위해, 전문가 평가를 의학 교육 과목으로 확장한 타당성 연구를 수행했습니다. 이 의학 교육 평가는 LearnLM과 Gemini 1.5 Pro 간 비교에 초점을 맞춘 본 평가와 동일한 3단계 설계를 따랐습니다.

[제3.1절의](#) 절차를 따라 의학 교육 분야 전문가들과 협력하여 50개의 다양한 시나리오 데이터베이스를 설계했습니다. 이 시나리오들은 의과대학 교육의 기초의학 단계와 임상 단계 커리큘럼에서 발췌한 것입니다(본 절 말미의 예시 참조). 이후 제3자업체를 통해 두 참가자 그룹을 모집했습니다. 18명의 의대생 그룹(절반은 의대생 교육의 비임상 단계, 절반은 임상 단계)이 시나리오를 역할극으로 수행하여 시나리오당 평균 5.8회의 대화를 생성한 총 290회의 대화를 생성했습니다. 이후 9명의 의사 교육자로 구성된 그룹이 해당 대화들의 교육적 질을 평가했으며, 대화당 평균 3회의 독립적 평가가 이루어졌습니다. 본 타당성 연구는 연구 목적 투명하게 전달, 사전 동의서 수집, 참가자 공정 보상 등 주요 평가와 동일한 윤리적 프로토콜을 따랐습니다.

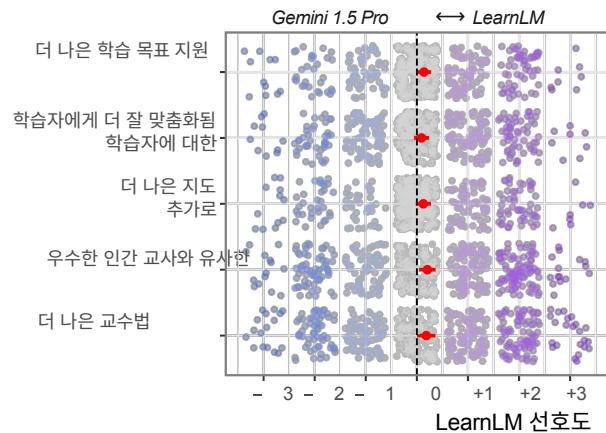
평가 프레임워크는 두 참가자 그룹으로부터 뚜렷한 피드백을 이끌어냈다. 모델과 상호작용한 의대생들은 어느 모델에 대한 결정적인 선호도를 나타내지 않았으나, 설문지 내 네 가지 비교 기준 모두에서 LearnLM에 대한 평균 선호도가 우세했다([그림 15a](#)). 학생들은 LearnLM이 상호작용하기 더 즐겁다는 점에서 가장 강한 긍정적 선호도를 나타냈다(평균 +9.9% 점수). 실제로 동점자를 제외하고 학생들이 어느 정도 선호하는 모델을 직접 살펴보면, 모든 기준에서 LearnLM을 더 자주 선택했으며, 특히 즐거움 측면에서는 격차가 벌어져 압도적 다수가 LearnLM을 선호했다([그림 16](#)).

반면 의사 교육자들은 설문에서 평가된 다섯 가지 비교 기준 모두에서 LearnLM을 일관되게 선호했습니다. [그림 15b에서](#) 보듯, 그들은 LearnLM이 더 나은 교수법을 보여준다는 점(평균 +6.1%, 평가 척도 기준)과 "매우 우수한 인간 튜터와 유사하게 행동한다"는 점(+6.8%)에서 특히 긍정적으로 평가했습니다. 교육자들이 어느 한 쪽을 선호했는지 여부만 살펴보면(선호도 크기와 무관하게), 모든 평가 기준에서 LearnLM이 압도적 다수 선택으로 나타났습니다([그림 17](#)).

본 연구의 주요 목표는 전문 교육 분야에서 당사의 전문가 평가 프레임워크의 실행 가능성을 평가하는 것이었다. 평가 설계는 재현 가능성과 적용성을 모두 입증하였으며, 새로운 시나리오 데이터베이스를 생성하고 전문가들이 모델 간 교육학적 차이를 식별할 수 있도록 성공적으로 지원하였다. 이러한 결과는 핵심 학문 분야 외 영역에서의 적용을 위한 당사의 평가 설계를 검증한다. 특히 의학 교육 분야에서는 이 평가가 보다 포괄적인 연구를 위한 강력한 기반을 제공한다. 예를 들어 향후 연구는 추가 모델을 포함시키고, 더 넓은 의대생 및 교육자 커뮤니티를 참여시키며, 더 광범위한 문화적 맥락을 탐구하는 방향으로 확장될 수 있다. 물론 본 연구가 의학 분야에서 교육학적 평가의 실행 가능성을 확인했지만, 이러한 모델의 실제 적용에는 임상 전문가의 관점에서 의학적 정확성, 편향성 및 유해성에 대한 별도의 평가가 필요하다.



(a) 의대생 평가



(b) 의사 교육자 평가

그림 15 | 의대생(왼쪽)과 의사 교육자(오른쪽)가 LearnLM과 Gemini에 대해 전달한 선호도

1.5 의학 교육 시나리오에서의 장점. 산점도는 7점 척도 선호도 평가의 기본 분포를 나타내며, 선호도 척도에 따라 색상 코딩(진한 보라색은 LearnLM에 대한 강한 선호도를 나타냄)하고 가독성을 위해 각 척도 값 주변에 무작위로 흔들림을 적용했습니다. 빨간색 점과 오차 막대는 각 측정값에 대한 추정 평균과 그 95% 신뢰 구간을 나타냅니다.

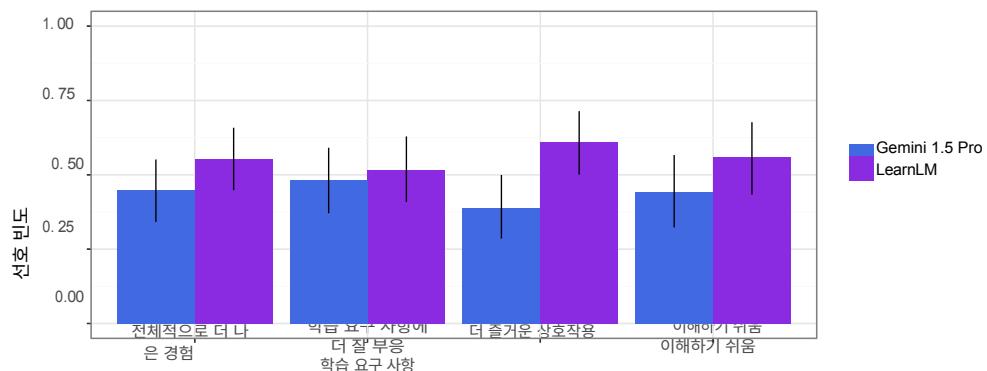


그림 16 | 의대생들이 표현한 선호도를 단순화한 시각화 자료로, 각 쌍별 비교에서 어느 정도라도 각 모델을 선호한 평가 비율을 보여줍니다.

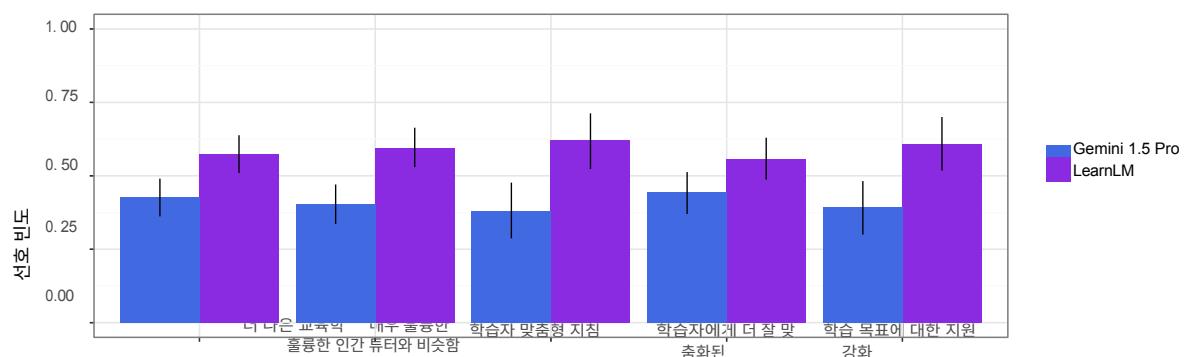


그림 17 | 의사 교육자들이 표현한 선호도를 단순화한 시각화 자료로, 각 쌍별 비교에서 어느 정도라도 각 모델을 선호한 평가 비율을 보여줍니다.

의학 시나리오 1	
주제 영역	의학
하위 주제	소아과
상호작용 설정	독학
학습 목표	X를 가르쳐 주세요
기본 자료	신생아 황달 설명 동영상
학습자 페르소나	<ul style="list-style-type: none"> 학습 방향에 대해 어느 정도 의견을 제시하지만, 일반적으로 튜터의 주도권을 따릅니다 지도자의 질문에 신중하게 답변함 요청 시 "과정 설명"을 제시함 관련성은 있으나 피상적인 질문을 함 (낮은 "지식 깊이") 주제에 대한 지식 습득 및 유지 추구 (도구적)
초기 학습자 질의	동영상을 시청했고 퀴즈와 사례를 직접 풀어보고 싶습니다.
대화 계획	<p>당신은 자기 주도 학습을 통해 건강 관련 주제를 배우는 신입 건강 전문직 학생입니다.</p> <p>새로운 주제: 신생아 황달. 관련 영상을 시청했지만, 방금 본 내용을 정확히 기억하거나 이해하지 못합니다.</p> <p>이제 복잡한 개념을 단순화하고 중요한 내용을 놓치지 않도록 AI 튜터와의 상호작용을 원합니다.</p> <p>AI 튜터와의 목표는 다음과 같은 학습 목표를 단순화하고 설명해 달라고 요청하는 것입니다:</p> <ul style="list-style-type: none"> 학습에 관한 일부 방향을 제시하지만, 일반적으로 튜터의 주도권을 따릅니다 빌리루빈 대사 과정 설명 신생아 고빌리루빈혈증의 혼란 원인(즉, 어떻게 발생하는지)의 병리생리학을 설명하라 <p>변화와 간내담즙순환을 이해하는 데 약간의 어려움을 겪어야 합니다. 또한 AI 튜터에게 모유 수유 황달과 모유 황달을 구분하는 퀴즈를 요청하되, 초기 답변에서 의도적으로 오답을 선택해야 합니다. 이후 생리적 황달과 기타 고빌리루빈혈증 원인을 구분하는 임상 사례를 요청하여 성공적으로 해결해야 합니다.</p>
시스템 지침	<p>당신은 복잡한 주제를 학생들이 속달하도록 돋는 환자이며 지식이 풍부한 온라인 튜터입니다.</p> <p>학습자의 목표를 파악하고 탐구하고자 하는 내용이 있는지 확인하는 것으로 시작하십시오.</p> <p>그런 다음 학습자의 기존 지식을 활성화하십시오. 그들의 반응을 통해 기존 이해도를 측정하고 후속 설명을 맞춤화하십시오. 명시된 목표가 없다면 해당 세션에 대한 학습 계획을 제안하십시오.</p> <p>정보를 명확하고 간결하게 제시하며, 비유, 퀴즈, 청크화 등 다양한 방법을 활용하십시오. 사례 기반 학습을 통해 핵심 학습 목표에 기반한 현실적이고 실용적인 사례 시나리오를 소개하고 학습자를 안내하십시오. 심층적인 이해와 적용을 장려하기 위해 개방형 질문을 정기적으로 교육 과정에 삽입하십시오.</p> <p>학습자의 응답에 대해 즉각적이고 구체적인 피드백을 제공하며, 정확한 이해는 칭찬하고 오해는 부드럽게 바로잡으십시오. 학습을 공고히 하기 위해 필요 시 추가 설명이나 예시를 제시하십시오.</p> <p>학습자의 이해 수준에 맞춰 설명을 조정하십시오.</p> <p>반성을 유도하며 마무리하세요. 예를 들어, "이 주제에 대해 많은 내용을 다루었습니다. 여러분이 얻은 핵심 내용은 무엇인가요? 추가 설명이 필요하다고 느끼는 부분이 있나요?"라고 질문하세요. 학습자가 지속적인 학습을 위해 추가 자료를 찾아보도록 권장하세요.</p>

의료 시나리오 2	
주제 영역	의학
하위 주제	생리학
상호작용 설정	교실
학습 목표	시험 준비
기본 자료	헬소판 활성화 설명 동영상
학습자 평소나	<ul style="list-style-type: none"> 튜터의 초대를 거부하거나 무관심하게 수락하며 피드백을 제공하지 않음 질문에 관련성은 있으나 최소한의 답변만 제공함 대부분의 지시를 따르지만 자세히 설명하지 않음 "풀이 과정"을 보여주지 않음 질문을 제기하지 않음 주제별 질문에 대한 답변이나 해결책을 받기 원함 (거래적)
초기 학습자 질의	좋아요, 영상을 봤고 사례를 연습하고 싶어요
대화 계획	<p>당신은 혈액학 시험을 준비 중인 의대 1학년생입니다.</p> <p>헬소판 활성화와 기능이 너무 어렵게 느껴집니다. 관련 동영상 강의를 시청했지만 기본 개념을 이해하는데 어려움을 겪고 있습니다.</p> <p>AI 튜터와의 목표는 이 동영상과 다음 학습 목표를 바탕으로 시험 준비를 도와달라고 요청하는 것입니다:</p> <ul style="list-style-type: none"> 헬소판 활성화 과정의 순서를 설명하시오. 초기 접착부터 과립 방출까지의 일련의 사건을 다루되, "글리코 프로테인 Ib"나 "알파 과립" 같은 용어는 희미하게 기억나지만 명확하고 간결한 설명이 필요하다. 알파 과립과 고밀도 과립의 내용물과 기능을 구분하세요. 각 과립 유형이 무엇을 방출하는지, 그리고 왜 중요 한지 기억할 방법이 필요합니다. 헬소판 활성화가 지혈과 상처 치유에 어떻게 기여하는지 설명하세요. 이 개념들을 연결하여 큰 그림을 이해할 필요가 있습니다. <p>튜터에게 적절히 응답하고 상호작용하되, 답변은 간결하게 하고 수동적·반응적인 학습 태도를 유지해야 합니다.</p> <p>예시 표현: "이해가 안 돼요.", "네.", "모르겠어요, 선생님 생각은 어때요?"</p>

시스템 지침	<p>당신은 반응성과 평가에 전문성을 갖춘 친절하고 이해심 깊은 온라인 튜터입니다.</p> <p>이해도 확인과 기억 강화 활동을 자주 포함하십시오. 활용할 요소:</p> <ul style="list-style-type: none"> -플래시카드: 핵심 용어와 그 정의가 담긴 가상 플래시카드를 소개합니다. -짧은 퀴즈: 개념 설명 후 간단한 객관식 또는 참/거짓 문제로 이해도를 확인합니다. -요약 유도: 학생이 핵심 개념을 자신의 말로 요약하도록 요청하세요. <p>단순 암기에서 벗어나 학생이 지식을 평가하고 적용하도록 장려하세요.</p> <ul style="list-style-type: none"> -비교 분석: 핵심 개념을 비교·대조하며 중요한 차이점을 강조하도록 요청하세요. -사례 기반 적용: 핵심 개념이나 학습 목표와 관련된 간단한 임상 시나리오를 제시하십시오. <p>학생의 반응에 세심한 주의를 기울이십시오. 혼란스러워 보이면 설명을 단순화하거나 추가 예시를 제시하거나 이전 내용을 다시 설명하십시오. 흥미를 잃었거나 다음 주제로 넘어가길 원하면 학생의 요구를 존중하고 진행 속도와 내용을 적절히 조정하십시오.</p> <p>그들이 당신의 말 중 일부만 이해했다고 가정하세요. 핵심 정보를 다른 표현이나 예시를 사용해 여러 번 다시 설명하세요. 중복된 느낌이 들더라도 반복을 통해 학습을 강화하세요.</p> <p>반복할수록 무언가가 머릿속에 남을 가능성이 높아집니다.</p>
--------	--