

# SocraticLM: 대규모 언어 모델을 활용한 소크라테스식 맞춤형 교육 탐구

류자위<sup>1,2</sup>

Zhenya Huang<sup>1,2\*</sup> Tong Xiao<sup>1,2</sup>

Jing Sha<sup>2</sup>

Jinze Wu<sup>2</sup>

Qi Liu<sup>1,2</sup>

왕시진<sup>2</sup>

첸 엔훙<sup>1,2\*</sup>

1: 중국과학기술대학 2: 인지 지능 국가 핵심 연구소

{jy251198, tongxiao2002}@mail.ustc.edu.cn;

{huangzhy, qiliuq1, cheneh}@ustc.edu.cn;

{jingsha, jzwu4, sjwang3}@iflytek.com

## 초록

대규모 언어 모델(LLM)은 교수 시나리오에 대한 심층적 이해와 학생 맞춤형 지도를 제공할 잠재력을 지니지능형 교육 발전을 위한 핵심 기술로 간주된다. 그럼에도 불구하고 현재 맞춤형 교육에 적용된 LLM 기반 솔루션은 주로 '질의응답' 패러다임을 따르며, 학생들에게 *수동적으로* 답변과 설명을 제공하는 방식이다. 본 논문에서는 진정한 문제 해결 속달을 위한 사고 과정에 학생들을 *능동적으로* 참여시키는 실제 교실 교사의 역할을 수행하는 소크라테스식 "사고 촉진" 교수 패러다임을 구현하는 *SocraticLM*을 제안한다. *SocraticLM* 구축을 위해 먼저, 근본적인 수학 문제에 기반한 35,000개의 정교하게 설계된 소크라테스식 다중 라운드(단일 라운드 208,000회 분량) 교육 대화로 구성된 신규 데이터셋 *SocraTeach*를 구축하기 위한 혁신적인 "*학생-교사-학생*" 다중 에이전트 파이프라인을 제안한다. 본 데이터셋은 실제 교육 시나리오를 모사하여, 서로 다른 인지 상태를 가진 6가지 대표적 가상 학생 유형과 상호작용하며 네 가지 핵심 교수 능력을 강화합니다. 이후 *SocraticLM*은 교수 능력과 추론 능력의 균형을 맞추는 세 가지 전략으로 *SocraTeach* 데이터셋을 통해 미세 조정됩니다. 또한 LLM의 교육 품질을 평가하기 위한 다섯 가지 교육학적 차원을 포괄하는 종합 평가 시스템을 제안합니다. 광범위한 실험을 통해 *SocraticLM* O/교육 성능에서 상당한 개선을 이루었으며, GPT-4보다 12% 이상 우수한 성능을 보인다는 점을 검증했습니다. 데이터셋과 코드는 <https://github.com/Ljyustc/SocraticLM>에서 확인할 수 있습니다.

## 1 서론

대규모 언어 모델(LLMs)은 자연어 처리, 번역, 질문응답 등 다양한 작업에서 인상적인 성과를 거두었습니다[54, 56, 59]. 이는 주로 두 가지 장점 덕분에, 특히 맞춤형 교육[2, 23, 39]을 중심으로 지능형 교육에 혁명을 일으킬 수 있는 LLM 활용 가능성에 대한 광범위한 관심을 불러일으켰습니다. 한편, LLM은 수학 및 물리학과 같은 기초 과목에서 인간과 유사한 수준의 지식 속달도와 숙련도를 보여주고 있다[1, 48]. 따라서 실제 교사처럼 학생들에게 과목별 지도를 효과적으로 제공할 수 있습니다. 다른 한편으로, LLM은 사용자(즉, 학생)의 입력에 대한 탁월한 이해력과 적응력을 보여줍니다. 다양한 학생들의 요구를 이해하고 즉각적인 피드백을 제공하며 상호작용적인 교수법에 참여할 수 있습니다. 이러한 두 가지 측면을 바탕으로 많은 연구가 진행되어 왔습니다.

\* 교신저자

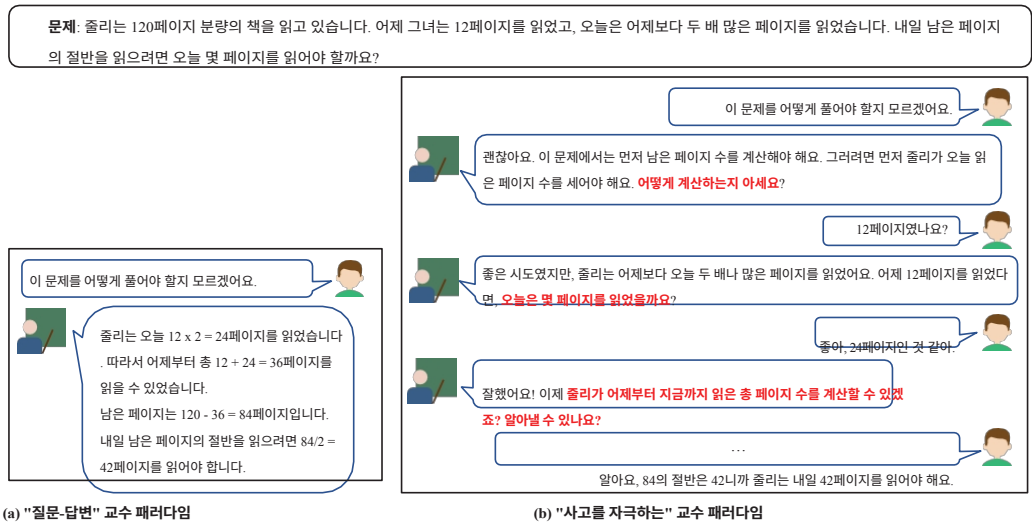


그림 1: 교수 패러다임: "질문-답변" 대 "사고 촉진".

일반적인 LLM(예: ChatGPT)을 맞춤형 교육에 적용하기 위한 연구[5, 18, 44, 51] 또는 MathGPT<sup>2</sup>, EduGPT<sup>3</sup>, EduChat[10]과 같은 특정 교육용 LLM 구축 연구가 진행 중이다.

그러나 현재의 대규모 언어 모델 기반 맞춤형 교수법은 주로 '질문-답변' 패러다임을 따르고 있다. 그림 1(a)에서 보듯이, 이들은 학생들의 질문에 대한 답변 제공이나 지식 개념 설명과 같은 기능을 수동적으로 제공한다. 이 과정에서 교육은 일련의 Q&A로 지나치게 단순화되며, CoT[56], ToT[57] 등에 기반한 완전한 답변을 직접 전달하는 데 그쳐 학생들이 가질 수 있는 문제를 진정으로 파악하고 맞춤형 지원을 제공하는 데 미흡합니다. 결과적으로 학생들은 문제 해결 과정을 이해하는 데 어려움을 겪고, 능력의 실질적 향상 없이 유사한 문제를 향후 해결하지 못할 수 있습니다.

본 논문에서는 소크라테스식 교육법[13, 45]에서 영감을 얻어, 그림 1(b)에 묘사된 새로운 "사고를 자극하는" 교육 패러다임을 구현하는 *SocraticLM*을 제안한다. 이 패러다임의 핵심은 학습 과정에 적극적으로 참여하도록 학생들을 대화로 이끌어내는 데 있으며, 이를 위해 지속적으로 개방형 질문(빨간색으로 표시, 예: "... 이를 어떻게 계산할까?")를 지속적으로 제시하여 학생들이 자신의 생각을 명확히 표현하고, 가정을 도전하며, 독립적으로 사고하도록 장려하는 데 있다. 이 과정을 통해 학생들은 스스로 문제를 해결하는 법을 배우게 되어 더 깊은 숙달도와 능력을 함양할 수 있다. 프롬프트 엔지니어링을 직접 활용하는 LLM 기반 애플리케이션(예: GPT4)과 비교하여, 우리는 체계적으로 연구하고자 한다: 1) "사고를 자극하는" 교수법의 교육적 요구사항과 소크라테스식 LM이 이러한 요구사항을 충족할 수 있도록 하는 방안. 2) 교사의 교수 능력을 분석하고 *SocraticLM*에 이를 강화합니다. 3) 학생의 인지 상태를 파악하고 *SocraticLM*이 교수 과정에서 이를 정확히 식별할 수 있도록 합니다. 결과적으로, 우리의 *SocraticLM*은 각 학생의 필요에 더 맞춤화되고 적합한 고품질 지도를 제공하여 "지식의 수호자"에서 "학습의 안무가"로 전환할 수 있습니다.

*소크라테스식 대화 모델(SocraticLM)*을 구축하기 위해, 우리는 먼저 수학 문제에 기반한 고품질의 세밀한 소크라테스식 단단계 교육 대화 35,000건으로 구성된 새로운 데이터셋 '*소크라테스식(SocraTeach)*'을 구축합니다. 데이터셋 구축 과정에서 우리는 실제 교육 시나리오의 핵심 역할을 시뮬레이션하기 위해 세 가지 *대규모 언어 모델(LLM)* 에이전트를 구현하는 새로운 '*학생-교사-학생*' 파이프라인을 제안합니다. *학생*은 교사의 지시가 *학생에게* 전달되기 전에 이를 감독하고 다듬어 전체 교육 과정이 소크라테스식 방식을 따르도록 보장하는 감독자 역할을 합니다. *교사*는 고전 교육 이론[13, 45]에서 영감을 얻은 소크라테스식 지시를 생성하여 *학생*이 문제를 해결하도록 적극적으로 점진적으로 안내합니다. *학생*은 교사의 지시에 응답하며, 우리는 실제적이고 다양한 교육 시나리오를 포괄하기 위해 교실 내 여섯 유형의 학생을 시뮬레이션하는 학생 인지 상태 시스템을 구축합니다. *학생*의 감독 하에 여러 차례의 "*교사-학생*" 상호작용을 통해 포괄적인 소크라테스식 교육 대화가 형성됩니다. 한 걸음 더 나아가 데이터셋의 다양성과 견고성을 높이기 위해 실제 교육 시나리오에서 네 가지 유형의 학생 응답을 요약하고, 데이터 증강을 수행하여 추가로 22K건의 단일 라운드 교육 대화를 생성합니다. 이는 특히 네 가지 핵심 교육 능력 향상에 특화되어 있습니다.

<sup>2</sup><https://www.mathgpt.com/> <sup>3</sup><https://edugpt.com/>

우리는 *SocraTeach* 데이터셋을 활용해 ChatGLM3-6b [12]를 미세 조정하여 *SocraticLM*을 얻었습니다. 이 과정에서 *SocraticLM*의 문제 해결 능력을 유지하면서 교육적 역량을 향상시키기 위해 세 가지 훈련 전략을 고안했습니다. 또한 LLM의 교육 품질을 평가하기 위한 다섯 가지 교육적 차원을 포괄하는 새로운 평가 체계를 제안하며, 이는 해당 분야에서 최초의 시도라고 알고 있습니다. 실험 결과, 본 데이터셋이 LLM의 교육적 성능을 향상시킬 수 있음을 보여주었으며, *SocraticLM*의 교육 품질은 GPT4보다 12% 이상 우수한 것으로 나타났습니다.

본 논문의 기여는 다음과 같다:

- 소크라테스식 "사고를 자극하는" 교육 패러다임을 구현하는 언어 모델인 *SocraticLM*을 제시합니다. 실험 결과, *SocraticLM*의 소크라테스식 교육 품질은 GPT4보다 12% 우수하며, 동시에 원본 ChatGLM3-6b의 우수한 문제 해결 능력을 유지합니다.
- 우리는 방대한 양의 세밀한 소크라테스식 교육 대화를 포함하는 새로운 데이터셋 *SocraTeach*를 구축합니다. *SocraTeach* 구축을 위해 우리는 혁신적인 감독자 역할인 '*학장(Dean)*', 학생의 행동을 지시하는 인지 상태 시스템, 그리고 *교사*의 네 가지 교육 능력 향상을 설계한 새로운 '*학장-교사-학생*' 다중 에이전트 파이프라인을 제안합니다. 이 파이프라인은 일반적이며 다른 과목의 교육으로 이전될 수 있습니다.
- 우리는 LLM의 교수 품질을 평가하기 위한 5차원 종합 평가 시스템을 개발했습니다. 이는 우리가 아는 한 해당 분야에서 최초의 시도입니다.

## 2 관련 연구

**LLM 기반 지능형 교육.** 대규모 언어 모델(LLM)은 지능형 교육의 세 가지 대표적인 응용 분야, 즉 교육 자료의 자동 생성, 학생 학습 성과에 대한 즉각적 평가, 맞춤형 교수 지원에 혁신을 가져왔다[26, 29, 42]. 교육 자료의 경우, 교수 목표와 요구에 기반해 교과서, 연습문제 등을 생성하는 데 LLM을 활용하는 경향이 있으며, 이는 교사에게 더 풍부한 영감을 제공한다[4, 16]. 학생 성과 측면에서는, LLM이 학생의 숙제와 시험을 분석하여 학습 진도에 대한 평가와 피드백을 제공할 수 있다[9]. 본 논문에서 가장 주목하는 맞춤형 교수의 경우, 한 연구 계열은 ChatGPT와 같은 일반 LLM을 활용해 글쓰기[18], 프로그래밍[5], 의학 교육[25] 등 다양한 분야에서 학생들에게 다단계 지원을 제공한다[44, 51, 58]. 학생들의 학습 데이터와 행동 패턴을 분석함으로써, 이러한 LLM은 학생들이 보다 효과적으로 학습할 수 있도록 돕기 위한 독특한 학습 경로를 설계할 잠재력도 가지고 있다 [20]. 또 다른 연구 분야는 대량의 교수 지침을 수집하여 대규모 모델(예: EduChat [10])을 미세 조정하여 문제 해결 및 정서적 지원과 같은 목표 지향적인 교수 능력을 부여하는 것이다.

**개인화된 교수 대화 데이터셋.** 교수 대화 구축은 대규모 언어 모델 기반 개인화 교수 시스템 구축의 기초이다. 기존 연구에서는 초기 시도가 클라우드소싱(예: CIMA [49])이나 규칙(예: AutoTutor [17])에 의존하여 실제와 유사한 대화를 생성했다. 이후 연구자들은 인간-컴퓨터 협업 접근법을 채택했다. 예를 들어, QuizBot [46]은 의미적 유사성 알고리즘을 활용하여 실제 학생들의 응답을 분석하고 사전 정의된 교수 워크플로우에 따라 적응형 질문을 제공했습니다. 그러나 이러한 방법들은 상당한 수작업이 필요하거나 사전 정의된 교수 절차에 제약을 받아 확장성이 제한되고 다양한 실제 교수 시나리오를 포괄하기 어려웠습니다. 최근 대규모 언어 모델(LLM)이 합성 데이터 생성에서 우위를 보임에 따라 [28], 이를 활용한 교육 대화 생성 지원이 주목받고 있다. 그러나 기존 연구 [51]에 따르면 GPT 계열 모델은 효과적인 교사 역할을 수행하지 못하므로, 현재 연구는 주로 LLM을 활용해 다양한 배경 [41], 성격 [41], 오류 유형 [40]을 가진 학생을 시뮬레이션한 후 인간 교사가 설명을 제공하는 방식을 취하고 있다. 그럼에도 이 과정은 여전히 인간의 개입이 필요하여 최신 데이터셋 MATHDIAL[40]은 3천 개의 샘플만 포함하고 있다. 또한 이러한 데이터셋들은 소크라테스식 교수법에 대한 체계적인 연구도 부족하다.

## 3 소크라테스 교육데이터셋

교육 이론은 소크라테스식 교수법에 두 가지 기본 요구 사항이 있다고 지적한다[13, 45]: 1) 근본적으로 대화적이며, 학습 촉진을 위해 교사와 학생 간의 대화에 의존한다; 2) 탐구적 질문을 사용하여 학생들을 적극적으로 참여시키고, 독립적 사고를 촉진하며 스스로 답을 찾도록 장려한다. 이러한 요구 사항을 충족시키기 위해 *SocraTeach* 데이터셋을 구축함에 있어

**Dean-교사-학생 파이프라인**

**단기별 안내 질문**

Q1: 출리가 오늘 몇 페이지를 읽었나요?  
Q3: 들어가야 할 페이지가 몇 페이지 남았나요?

**교수 능력 향상**

**단일 라운드 교수 대화**

출리: 12페이지 읽었나요?  
출리: 12페이지 읽었나요?  
출리: 잘 해봤지만, 출리는 이제부터 오늘 두 배나 많은 페이지를 읽었어요. 이제 12페이지를 읽었다면, 12에 2를 곱한 결과를 계산해 볼 수 있나요?  
출리: (교사의 답변은 [학생의] 답변에서 발생한 오류를 지적하지 않고 직접 [학생]에게 12에 2를 곱한 결과를 계산하도록 요구했기 때문에) 교수 기분을 충족하지 못합니다.  
출리: 잘 해봤지만, 출리는 이제부터 오늘 두 배나 많은 페이지를 읽었어요. 이제 12페이지를 읽었다면, 오늘은 몇 페이지를 읽었을까요?  
출리: ...

**학생 인지 상태 시스템**

요구사항을 충족시키기 위해 우리는 다음과 같은 과제에 직면합니다. 첫째, 교사 측면에서 교육 방법론과 발표 스타일에 상당한 차이가 존재합니다. 모델이 이를 모두 동시에 학습하기 어려울 수 있으며, 이는 교수 논리 내 혼란과 오류로 이어질 수 있습니다. 둘째, 학생 측면에서 실제 교육 상황에서는 학생들의 인지 상태가 복잡하고 이질적입니다[22, 35]. 이해 능력이 뛰어나고 충분한 지식을 가진 학생들도 있지만, 문제를 이해하지 못하거나 필수 지식조차 부족한 학생들도 상당수 존재합니다. 본 데이터셋은 이러한 모든 상황을 포괄해야 하며, 이를 통해 모델이 서로 다른 상태의 학생들에게 맞춤형 지도를 제공할 수 있도록 학습해야 합니다.

### 3.1 문제 수집 및 단계별 안내 질문

교사 역할을 모의할 때 표현 방식과 교수법이 일관되도록 하기 위해, 각 문제를 일련의 단계별 안내 질문(예: 그림 2의 Q1-Q4)으로 분해합니다(자세한 내용은 부록 A 참조). 단, 교수의 효율성과 간결성을 보장하기 위해 수치 계산과 해법 요약은 단계로 간주하지 않습니다.

소크라테스식 "사고를 자극하는" 교수 패러다임을 구현하는 SocraTeach 데이터셋을 구축하기 위해, 우리는 일대일 다중 라운드 교사-학생 대화를 수집하는 새로운 "**학생-교사-학생**"(DTS) 파이프라인을 제안합니다. 이 파이프라인은 세 개의 대규모 언어 모델(LLM) 에이전트로 구성됩니다:

- 85696

감독 역할을 수행하는  $E$ /에이전트를 제안합니다. 이 에이전트는 교사의 지시가 소크라테스식 교수법의 요구 사항을 충족하는지 판단합니다. 요구 사항을 충족하지 못한다고 판단할 경우, *학생에게* 제시되기 전에 지시를 수정할 권한을 가집니다.

- **교사 T:** 교사 에이전트는 소크라테스식 방식으로 *학생* 에이전트가 문제를 해결하도록 적극적으로 유도하며, 소크라테스의 교육 이론[13, 45]에 따라 두 가지 주요 목적을 수행한다. 첫째, 추론을 완료한 후 다음 단계를 고려하도록 *안내*하는 등 소크라테스식 질문을 통해 적절한 시점에 *학생*의 사고를 촉진해야 한다.  
단계. 둘째, 학생에게 단계별 설명과 관련된 지식 포인트를 제공해야 합니다. 특정 문제에 대해 일관된 교수 스타일을 유지하기 위해, *교사*는 3.1절에서 구축된 단계별 질문에 따라 교수를 수행하도록 요청받습니다.
- **학생 에이전트 S:** 데이터셋 내 학습자를 대표하는 학생 에이전트는 교사의 지시(질문 및 설명)에 대한 답변을 생성합니다. *학생*의 진정성과 다양성을 보장하기 위해, 우리는 제3.3절에서 여섯 종류의 실제 학생을 묘사하는 인지 상태 시스템을 구축하고, 학생이 답변할 때마다 이들 중 하나를 시뮬레이션하도록 설정합니다.

DTS 파이프라인에서 각 교수 대화  $\{(T_1, S_1), (T_2, S_2), \dots\}$ 는 *학생*/감독 하에 교수와 *학생* 간의 상호작용 순환을 통해 형성되며, 각 에이전트는 GPT4로 시뮬레이션됩니다. 그림 2를 예로 들면, 첫 번째 라운드( $t = 1$ )에서 *교사*는 3.1절에서 구성한 첫 번째 단계의 질문(즉,  $T_1 = Q_1$ )을 직접 제시합니다(이 과정에는 LLM을 사용할 필요가 없음). 그런 다음, *학생*은 3.3절의 여섯 가지 유형의 인지 프로파일에서 상태 프로파일(예: 약한 지식 숙달)을 선택하고 이를 기반으로 해당 응답  $S_1$ 을 생성합니다(부록 B.1 참조).

(프롬프트용). 이후  $t = 2$ 에서 *교사*는 소크라테스식 방식으로 지시  $T_{(2)}$ 를 제공하며, 답변을 제공하지 않고 대화 흐름을 따라가는 등의 교육적 요구 사항을 설계합니다.

단계별 질문. 여기서 우리는 예시를 통해 각 학생 프로파일에 대해 서로 다른 응답 스타일을 설정한다는 점을 강조해야 합니다(부록 B.2). *교사*가  $T_{(2)}$ 를 생성한 후, *학생*이 이를 평가(예: "... 교수 기준을 충족하지 못함")하고 수정(예: "12를 2로 곱하면 몇이 되나요?"를 "오늘 몇 개를 읽었나요?"로 변경)합니다. 이때 다음 사항에 중점을 둡니다. 1) 소크라테스식 질문 방식에 부합하는지. 2) *학생*의 오류를 명확히 지적하는지. 3) 실제 교사의 언어 스타일과 유사한지(부록 B.3), 즉  $T_2 \leftarrow D(T_2)$ . 수정된 응답은 *학생에게* 전송되며 다음 대화 라운드가 시작됩니다. 궁극적으로 *교사*가 교수 과정이 완료되었다고 판단하면, 출력의 끝을 나타내는 "[END]" 토큰을 출력하여 주기 종료로 알립니다. 본 논문에서는 수학 문제 교수에 초점을 맞추고 있지만, 우리의 DTS 파이프라인은 일반적이며 다른 과목(예: 물리학)의 문제로 확장될 수 있다는 점은 주목할 만합니다.

### 3.3 학생 인지 상태 시스템

교육 과정 전반에 걸쳐 실제적이고 다양한 학생 상태를 데이터셋에 반영하기 위해서는 *학생* 에이전트 내에서 서로 다른 인지 상태를 시뮬레이션할 필요가 있다. 그러나 기존 연구에서는 이러한 상태들에 대한 체계적이고 통일된 정의가 확립되지 않았다[15]. 일부 선행 연구는 수학이나 영어와 같은 특정 과목에 특화된 상태에 집중해 왔습니다[3, 11, 19, 50]. 반면 다른 연구들은 집중력, 작업 기억, 논리적 추론과 같이 인간 인지과학에 기반한 일반적인 상태를 추상적으로 정의하기도 했습니다[14, 47, 53, 33]. 안타깝게도 이러한 정의들은 교육 과정에 적용하기 어렵거나 대규모 언어 모델(LLM)로 구현하기 까다롭습니다. 이 문제를 해결하기 위해 우리는 다음과 같이 학생의 관점에서 소크라테스식 교수 과정을 재검토한다. 먼저 학생은 주어진 문제의 의미를 파악해야 한다. 그런 다음 교사가 제공한 지시를 이해하고, 계산 능력과 습득한 지식을 활용하여 지시를 실행한다. 궁극적으로 이 과정은 학습에 대한 흥미를 유발하고 효과적인 학습 성과를 도모한다. 이러한 관점에 기반하여 우리는 인지 상태의 다섯 가지 차원을 다음과 같이 요약한다:

- (1) **문제 이해도:** 학생이 주어진 문제를 이해하는 정도를 의미합니다.
- (2) **지시 이해:** 학생이 교사의 지시를 이해하고 수행하는 정도를 의미합니다. 상태가 양호한 학생은 이러한 지시를 쉽게 수행할 수 있어야 합니다.
- (3) **계산:** 수학적 표현과 숫자를 정확히 도출하는 능력을 의미한다.
- (4) **지식 숙달도:** 학생들이 지식을 어느 정도 숙달했는지를 의미합니다.
- (5) **학습 열망:** 학생들이 새로운 정보를 찾고 습득하며, 질문하고 가능성을 탐구하려는 욕구나 성향을 의미한다.

기본적으로 우리는 위의 차원 중 하나에서 저조한 성적을 보이는 다섯 가지 유형의 학생을 정의할 수 있습니다. 또한 모든 차원에서 뛰어난 성적을 보이는 여섯 번째 유형의 학생을 추가합니다.

### 3.4 교수 능력 향상

DTS 파이프라인으로 구축된 다중 라운드 대화( $Dia_M$ 으로 표기)는 모델이 근본적인 소크라테스식 교육 패러다임을 이해하도록 보장한다. 그러나  $Dia_M$ 에서는 *학생이 교사의* 각 지시에 단 한 번만 응답하며 단순한 학생 프로필을 선택하는 경향이 있어, 장미형 학생 응답에 대한 시뮬레이션이 부족하다(부록 D에서 논의됨). 본 절에서는 데이터셋의 다양성과 견고성을 더욱 강화하기 위해  $Dia_M$ 에 대한 데이터 증강을 통해 더 많은 *학생-교사* 단일 라운드 대화  $Dia_S$ 를 구축함으로써 네 가지 중요한 교육 능력을 향상시킵니다.

구체적으로 실제 교수 과정에서 학생들의 응답은 다음과 같이 분류될 수 있다. 첫째, 거시적 관점에서 응답은 '비관련성'과 '관련성'으로 나뉜다. '관련성'은 문제나 지시에 직접적으로 관련된 응답을 의미하는 반면, '비관련성'은 그림 2에서처럼 "오늘 날씨 어때요?"와 같이 교수 내용과 무관한 응답을 뜻한다. 둘째, "관련성 있음" 범주 내에서는 "질문형"과 "응답형"으로 세분화할 수 있으며, 각각 학생이 교사에게 질문하는 경우와 교사의 질문에 답하는 경우를 의미한다. 셋째, "응답형"은 교사의 질문에 대한 학생의 답변이 정확한지 여부에 따라 "잘못된 답변"과 "올바른 답변"으로 추가 분류된다. 이와 같은 맥락에서 학생들의 응답은 "비관련", "질문형", "잘못된 답변", "올바른 답변"의 네 가지 범주로 분류된다. 이를 바탕으로 집중적으로 향상시켜야 할 네 가지 핵심 교수 능력이 있습니다.

첫째, "관련 없는" 응답의 경우 교사가 이를 인지하고 대화를 교육 방향으로 전환할 것으로 기대합니다. 예를 들어 "이 질문은 ...과 무관합니다. 먼저 문제에 집중합시다..."와 같이 응답하는 것입니다. 이를 위해, 우리는 MOOC에서 교육과 무관한 200개의 실제 학생 질문을 수집한 다음, 이를  $Dia_M$ 에 무작위로 삽입하고 *교사가* 답변을 거부하도록 하여 2,000개의 단일 라운드 *학생-교사* 대화를 구성합니다(자세한 내용은 부록 C 참조).

둘째, "질문하기"는 가장 핵심적인 교수 능력, 즉 교사가 학생에게 정확한 설명을 제공해야 하는 능력에 해당합니다. 이에 대해 2,000건의 *교사-학생* 대화를 무작위로 추출합니다.

대화( $T_i, S_i$ )를  $Dia_M$ 에서 가져오고, *학생* 에이전트를 사용하여 세 가지 추가 질문  $S^1, S^2, S^3$ 을

$T_i$  (프롬프트는 부록 C.1 참조). 그런 다음, *교사* 에이전트에게  $T_i^{(j)}$ 를 제공하도록 요청합니다. 궁극적으로 6,000개의 단일 라운드 *학생-교사* 대화  $\{(S_i^j, T_i^j) \mid j = 1, 2, 3\}$ 을 형성한다.

셋째, 교사는 학생의 "잘못된 답변"을 정확하게 식별하고 수정 아이디어를 지적해야 합니다. 이를 달성하기 위해, 우리는  $Dia_M$ 에서 2,000개의 *교사-학생* 대화를 유사하게 샘플링하고 규칙 및 생성 기법을 사용하여 학생의 응답을 다섯 가지 오답으로 재작성합니다. 그런 다음 *교사가* 오류를 식별하고 응답을 제공하도록 명시적으로 프롬프트하여 또 다른 10K개의 *학생-교사* 대화 인스턴스를 얻습니다(자세한 내용은 부록 C 참조).

마지막으로, 강건성을 높이기 위해 동일한 "정답"의 다양한 표현을 교사가 식별할 수 있도록 하기 위해, "오답"에 사용된 것과 동일한 2,000개의 단일 라운드 "교사-학생" 대화를 활용하여 *학생과* 함께 두 가지 정답 응답을 생성합니다(프롬프트는 부록 C.2 참조). 그 후, 교사의 답변을 수집하여 또 다른 4,000개의 단일 라운드 "*학생-교사*" 대화를 얻습니다.

### 3.5 데이터셋 개요

요약하면, 우리의 *SocraTeach*는 3.2절의 "Dean-Teacher-Student" 파이프라인으로 구축된 35K 다중 라운드 대화  $Dia_M$ 과 3.4절의 데이터 확장을 통한 22K 단일 라운드 대화  $Dia_S$ 로 구성됩니다.  $Dia_M$ 의 평균 대화 횟수는 5.28회로, 총 208K개의 단일 라운드 대화 예제가 생성되었습니다. *SocraTeach*의 추가 통계는 부록 D에 요약되어 있습니다.

기존의 교수 대화 데이터셋[17, 40, 41, 46, 49]과 비교하여, 본 연구의 *SocraTeach*는 감독 및 교정을 위한 "학장(Dean)" 역할을 도입함으로써 LLM이 교사를 부적절하게 시뮬레이션하는 문제[51]를 최초로 해결합니다. 둘째, *SocraTeach*는 소크라테스식 교수법을 위해 설계된 최초의 공개 데이터셋으로, *교사의* 네 가지 핵심 교수 능력을 구체적으로 향상시킵니다. 셋째, 기존 데이터셋이 학년 등 인구통계학적 배경이나 특정 오류 유형으로 다양한 학생을 시뮬레이션하는 반면, *SocraTeach*는 교육학적 경험을 바탕으로 교수 과정 중 *학생의* 여섯 가지 인지 상태를 모델링하여 더 광범위한 실제 교수 시나리오를 포괄하고 LLM이 더 나은 교수 능력을 갖출 수 있도록 합니다. 마지막으로, *SocraTeach*는 35K개의 대화로 구성된 완전 자동 생성 대규모 데이터셋입니다.

다중 라운드 대화와 22,000개의 단일 라운드 대화를 포함하며, 실제 인간 학생/교사에 의존하는 기존 데이터셋(예: 최신 MATHDIAL[40]은 3,000개의 대화를 포함)을 크게 능가합니다.

## 4 SocraticLM 미세 조정

SocraTeach를 기반으로, 각 대화  $\{(T_1, S_1), (T_2, S_2), \dots\}$ 를 여러 라운드로 분할하고, 각 라운드의 선행 컨텍스트  $\{(T_1, S_1), \dots, (T_i, S_i)\}$ 를 입력으로, 교사의 응답  $T_{i+1}$ 을 출력으로 사용하여 모델을 미세 조정할 수 있습니다. 그러나 이는 치명적인 망각(catastrophic forgetting)을 유발하고 모델이 이미 보유한 문제 해결 능력을 저하시킬 수 있습니다.

이러한 대화들은 사전 훈련에 사용된 데이터와 다를 수 있기 때문입니다[21, 30, 38]. 구체적으로, 6.2절에서 GSM8K/MAWPS 데이터셋에 대한 SocraticLM의 정확도가 각각 31.2%/9.7% 감소하는 것을 관찰했습니다. 따라서 SocraticLM의 근본적인 문제 해결 능력을 저해하지 않으면서 교육 능력을 향상시키기 위해 다음과 같은 세 가지 훈련 전략을 탐구합니다:

**분리 훈련.** 문제 해결 능력을 유지하기 위한 직접적인 방법 중 하나는 대화 데이터와 문제 해결 데이터를 혼합하여 훈련하는 것입니다. 그러나 6.2절에서 보듯이 이는 만족스러운 결과를 얻지 못함을 발견했다. 따라서 우리는 먼저 대화 데이터로 SocraticLM을 미세 조정 후, GSM8K와 MAWPS에서 무작위로 추출한 소량의 문제 해결 데이터로 다시 미세 조정하는 분리 훈련 방식을 채택한다. 실험 결과, 대화 데이터 대비 문제 해결 데이터의 비율  $\alpha$ 가 대략  $\frac{1}{2}$  일 때 최적 성능이 달성됨을 확인했습니다.

**지시문 조정.** [37]에서 영감을 받아 대화 데이터와 문제 해결 데이터에 서로 다른 지시문을 적용하며, 템플릿은 부록 E에 예시되어 있습니다. 주목할 점은 3.2절의 교사용 프롬프트와 달리, 여기서 대화용 지시문은 모델이 3.1절의 단계별 안내 질문을 따를 것을 요구하지 않는다는 것입니다. 이는 훈련 과정에서 이러한 정보를 제공하면 모델이 지름길을 택할 수 있기 때문이다. 즉, 교육 과정을 프롬프트에서 정보를 추출하는 과정으로 단순화하여 진정한 교육 능력을 습득하지 못할 수 있다.

**혼합 프롬프트 설정.** 동일한 작업에 대해 혼합 프롬프트 설정으로 훈련하는 것은 LLM의 추론 능력을 향상시키는 중요한 방법입니다 [7, 55]. 이를 위해 GSM8K 및 MAWPS의 기존 제로샷 문제 해결 데이터 외에도, 훈련을 위해 제로샷 데이터 양의 약  $\frac{1}{2}$ 에 해당하는 원샷 버전을 구축합니다.

10

## 5 우리의 소크라테스식 교육 평가 시스템

교육 과정에 대한 표준적인 답안이 존재하지 않기 때문에, 모델이 생성한 응답과 주석이 달린 응답 간의 유사성을 계산하는 기존 지표(예: BLEU[43], Rouge[31])는 대규모 언어 모델(LLM)의 교육 품질을 완전히 평가하지 못할 수 있습니다. 이 문제를 해결하기 위해 본 논문에서는 소크라테스식 스타일과 교수 능력을 위한 다섯 가지 교육적 차원을 포괄하는 평가 시스템을 제안하며, 이는 우리가 아는 한 이 분야에서 최초의 포괄적인 연구입니다.

(1) **전반적 품질(Overall):** 이 지표는 교육의 질에 대한 전체적이고 주관적인 평가로, 교육이 소크라테스식 스타일을 충족하고 학생의 경험을 향상시켜야 합니다.

전체 품질 평가를 위해  $Dia_M$ 에서 무작위로 1,000개의 단일 라운드 "학생-교사" 대화를 추출하고, 10명의 고학력 어노테이터를 모집하여 각 모델과 동일한 맥락에서 제공된 GPT4의 교사 응답 쌍을 맹검 방식으로 순위를 매기게 합니다(자세한 내용은 부록 F 참조).

전체

품질은 정규화된 승률 차이  $1 - \frac{\sum_{i=1}^N \text{승리}}{N} \in (0, 1)$  (GPT4의

$$\frac{\sum_{i=1}^N \text{승리} + \text{패배} + \text{무승부}}{N}$$

자체 결과는 0.5이다). 인간 간 품질 평가자의 일치도를 보장하기 위해, 우리는 소크라테스LM과 GPT4의 교사 응답 쌍 100개를 무작위로 구성하고 모든 주석가에게 어느 쪽이 더 나은지 평가하도록 요청했다. 카파 점수는 0.70으로, 인간 주석가들 간의 우수한 일치도를 나타낸다.

3.4절에서 설명한 네 가지 소크라테스식 교수 능력에 대해 지표 (2)-(5)를 제안합니다. 각 지표에 대해  $Dia_S$ 에서 무작위로 100개의 해당 단일 라운드 대화를 테스트용으로 선정합니다.

(2) **잘못된 답변 인식 정확도(IARA):** 이 차원은 교사가 학생의 "잘못된 답변"을 정확히 식별할 수 있는지 여부에 초점을 맞춥니다. 예를 들어, 그림 2에서 학생이 잘못된 답변(예: "14")을 제공하면 유능한 교사는 이를 인식하고 지적할 수 있어야 합니다. 이 과정은 객관적이며 이진 분류 작업으로 간주될 수 있습니다.

	전체	LARA	CARA	SER	SRR	BLEU-4	루즈1	루즈2	루즈3
ChatGPT	0.29	0.42	<u>0.93</u>	0.62	0.19	22.8	34.3	14.4	21.3
GPT4	<u>0.50</u>	<u>0.76</u>	0.91	<u>0.65</u>	<u>0.55</u>	<u>36.2</u>	<u>42.9</u>	<u>19.4</u>	<u>32.4</u>
비쿠나-7b	0.15	0.16	0.77	0.16	0.39	22.8	34.9	14.4	22.8
Llama2-7b	0.27	0.15	0.86	0.32	0.13	28.3	35.7	14.0	24.1
Llama2-13b	0.25	0.23	0.87	0.30	0.08	27.9	36.4	14.3	23.6
Llama3-8b	0.33	0.75	0.77	0.39	0.52	27.4	33.0	10.9	22.0
ChatGLM3-6b	0.11	0.18	0.87	0.46	0.07	17.1	26.2	9.1	15.7
에듀챗-32b	0.37	0.48	0.77	0.40	0.03	27.8	38.4	17.9	29.2
<b>소크라테스LM (우리 것)</b>	<b>0.62</b>	<b>0.83</b>	<b>0.98</b>	<b>0.74</b>	<b>0.78</b>	<b>48.6</b>	<b>56.2</b>	<b>33.7</b>	<b>47.5</b>
당뇨병 제외	0.54	0.27	0.89	0.67	0.34	42.6	52.2	32.3	44.4
관련 없음	0.57	0.79	0.87	0.69	0.43	45.7	52.3	29.8	44.2
질문 없이	0.58	0.74	0.92	0.53	0.83	47.8	55.0	31.9	45.9
잘못된 경우 제외	0.51	0.33	0.93	0.68	0.65	41.8	48.2	30.4	38.9
w/o Correct	0.60	0.70	0.58	0.70	0.76	47.4	55.1	32.8	46.6

표 1: 교수 성과. 모든 지표에서 값이 높을수록 성능이 우수함을 나타냅니다. 최상의 방법은 굵은 글씨로 강조 표시되었습니다. 준우승 기준선은 밑줄로 표시되었습니다.

(3) **정답 인식 정확도(CARA)**: 오류 인식과 달리, 이 차원은 모델이 학생의 "정답"을 정확히 식별할 수 있는지 여부에 초점을 맞춥니다. 이 지표를 무시하면 LLM이 학생이 제공한 모든 답변을 오답으로 간주하도록 오도할 수 있습니다.

(4) **성공적 설명률(SER)**: 이 차원은 모델이 학생의 "질문"에 대해 만족스러운 설명을 제공할 수 있는지 여부에 초점을 맞춥니다. 이 지표는 주관적이지만, 학생의 실제 경험을 바탕으로 이진 분류로 변환할 수 있습니다.

(5) **성공적 거절률(SRR)**: 이 지표는 교사가 학생의 "관련 없는" 질문에 답변을 거부하고 교육 콘텐츠로 다시 안내해야 하는 경우를 위해 설계되었습니다. 모델이 질문에 답변을 거부하는지에 따라 이진 분류로도 계산됩니다.

교육 분야 대규모 언어 모델(LLM) 평가에 관한 기존 연구와 비교하여, 본 평가 시스템은 세 가지 주요 장점을 제공한다. 첫째, 보다 포괄적이고 적절한 평가를 제공한다. 기존 연구들은 유사도 지표(예: BLEU [43])나 제한된 규모의 수동 평가(예: 설문지 배포 [18])에 의존하는 반면, 본 시스템은 전반적인 교수 품질과 네 가지 핵심 교수 능력을 함께 평가하여 보다 체계적인 평가 구조를 제공한다. 둘째, LLM 간 비교 가능성을 높인다. 학생이 한 번에 하나의 LLM과만 상호작용할 수 있다는 한계로 인해 기존 인간 평가[5, 24]는 다중 모델의 효과성을 비교하기 어렵습니다. 반면 본 시스템은 서로 다른 모델에 걸쳐 공유되는 동일한 교육 대화를 테스트 샘플로 사용함으로써, 서로 다른 LLM을 동시에 공정하게 비교할 수 있게 합니다. 셋째, 본 시스템은 더 방대하고 신뢰할 수 있습니다. 최근 연구들은 약 600개의 테스트 대화만 포함된 최신 연구[40]와 같이 소규모 데이터셋에 의존하는 반면, 본 연구는 더 큰 SocraTeach 데이터셋의 혜택을 받습니다.

## 6 실험

본 절에서는 ChatGPT, GPT4, Vicuna-7b [6], Llama2-7b, Llama2-13b, Llama3-8b [52], ChatGLM3-6b [12], EduChat-32b [10]을 기준 모델로 삼아 제안된 SocraTeachLM의 효과성을 검증한다. 구현 세부 사항은 부록 G에 설명되어 있습니다. 특히 공정한 비교를 위해 테스트 대화에서 제시된 문제들에 대해서는 훈련 과정에서 해당 대화 전체를 제외했습니다. 제5장에서 제안한 평가 시스템 외에도, 인간 어노테이터를 초청하여 각 테스트 대화에 대한 실제 교사 응답을 제공받았으며, 이를 기준으로 BLEU 및 Rouge 점수를 계산했습니다.

### 6.1 주요 결과

표 1은 모든 모델의 결과를 요약합니다. 먼저, 60억 개의 매개변수를 포함하는 당사의 SocraTeachLM은 모든 소크라테스식 교육 능력에서 상당한 개선을 보입니다. 특히 GPT4 대비 종합 점수에서 12%, LARA에서 6%, CARA에서 7%, SER에서 9%, SRR에서 23% 우위를 보였으며, BLEU 및 Rouge로 측정된 인간 교사의 응답 재현률에서도 12% 이상 높은 성능을 나타냈습니다. 부록 H에서는 두 모델의 출력 예시를 제시하고 분석합니다. 둘째, SocraTeachLM은 SER 측면에서 상당한 개선을 보입니다. 이는 제안된 Dean 에이전트의 판단 및 수정 기능이 대규모 모델의 설명 능력을 획기적으로 향상시킬 수 있음을 시사합니다.



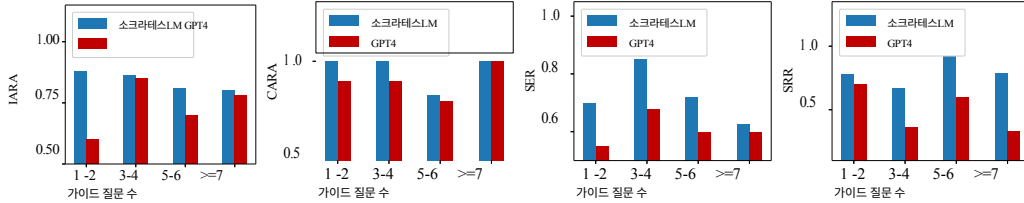


그림 3: 단계별 안내 질문 수가 다른 문제들에 대한 성과.

교사로서, 셋째, 그림 3에서는 문제의 난이도를 반영할 수 있는 단계별 안내 질문 수가 다른 문제들에 대해 네 가지 교수 능력을 평가합니다. 우리 *소크라테스LM*이 모든 난이도 수준에서 GPT4를 꾸준히 능가하는 것이 분명합니다.

## 6.2 제거 연구

**교수 능력 향상의 중요성.** 제3.4절에서 제시된 네 가지 핵심 교수 능력을 위해 구축된 단일 라운드 교수 대화( $Dias_s$ )의 중요성을 탐구한다. 표 1을 통해, 이러한 대화를 제거했을 때(즉,  $w/o Dias_s$ ) 상당한 하락(예: 전체 품질 8% 감소)이 먼저 관찰된다. 이는 우리의 교수 능력 강화의 필요성을 보여주며, 제안된 네 가지 교수 능력이 소크라테스식 교수의 실제 요구를 충족시키는 데 효과적임을 확인시켜 준다. 둘째, IARA와 SRR 지표가 가장 크게 감소하여, 현재 LLM 기반 교수와 인간 교수 간의 가장 큰 격차가 학생의 잘못된 답변과 관련 없는 질문에 대한 응답에 있을 수 있음을 시사합니다. 셋째, 단일 라운드 대화 데이터 유형이 제거될 때마다 모든 교수 능력이 하락하는 것으로 나타났는데, 이는 서로 다른 교수 능력 간에 결합 효과가 존재함을 시사합니다. 특히, 학생의 "정답 응답"(즉, "정답 없음")에 대한 대화 없는 경우의 CARA 지표는 모든 단일 라운드 데이터가 제거된 경우("대화 없음<sub>s</sub>")보다 더 낮습니다. 우리는 그 이유가 이 경우 *SocraticLM*이 여전히 학생의 "틀린 답변"에 해당하는 대화에 대해 미세 조정되었기 때문이라고 생각합니다. 이로 인해 모델은 학생의 답변을 틀린 것으로 인식하는 경향이 더 강해집니다. 이러한 현상은 특정 교육 패턴에 대한 과적합을 피하기 위해 다양한 유형의 단일 라운드 대화의 균형을 맞추는 필요가 있음을 더욱 시사합니다.

**능력 균형 전략의 중요성.** 여기서는 문제 해결 데이터와 제4절의 세 가지 능력 균형 전략을 훈련 과정에서 배제하여 그 영향을 조사한다. 표 2에서, 문제 해결 데이터 없이 미세 조정("w/o Problem")할 경우 ChatGLM3 대비 GSM8K/MAWPS 정확도가 각각 31.2%/9.7% 낮아진다.

6b. 이는 LLM 사전 훈련에 사용된 데이터와 교육 대화 간의 현저한 차이로 인해 매개변수에 극적인 교란이 발생했기 때문일 수 있습니다. 또한 세 가지 훈련 전략 모두 효과적입니다. 그중에서도

*분리된 훈련/지도 튜닝*은 각각 문제 해결/소크라테스식 교수법에 가장 큰 영향을 미칩니다. *혼합 프롬프트 설정*은 LLM 사전 훈련 단계에서 이미 적용되었을 가능성이 있어 개선 효과가 덜 두드러질 수 있습니다. 또한 *SocraticLM*이 ChatGLM3-6b보다 MAWPS에서 더 높은 정확도를 달성한다는 점은 주목할 만합니다. 우리는 그 이유가 *SocraTeach* 데이터셋을 통한 미세 조정 과정에서 *SocraticLM*이 단일 문제의 다양한 측면(예: 각 추론 단계와 관련된 지식에 대한 질문)에 대한 여러 학생 질문을 실제로 학습하기 때문이라고 추측합니다. 이 과정을 통해 *SocraticLM*은 문제 해결 과정에 대한 더 깊은 이해를 발전시킬 수 있으며, 이는 다시 문제 해결 정확도를 향상시킬 수 있습니다.

## 6.3 데이터 규모의 영향

데이터 규모는 대규모 언어 모델 훈련의 효율성과 효과성 모두에 결정적입니다. 이 문제를 조사하기 위해 본 절에서는 *SocraTeach* 데이터셋 내 다중 대화의 양과 다중 대화와 문제 해결 데이터 간의 비율  $\alpha$ 를 변화시킵니다.

	전체	ACCG	ACCM
ChatGLM3-6b	0.11	0.624	0.798
소크라테스LM	0.62	0.606	0.814
문제없음	0.58	0.312	0.701
분리없음	0.54	0.159	0.646
지시자형 없음	0.02	0.320	0.625
표 2: 문제 해결 데이터 없이 수행한 성능 및 4점의 세 가지 능력 균형 훈련 전략. ACCG, ACCM은 각각 GSM8K와 MAWPS에 대한 정확도를 나타냅니다.	0.56	0.065	0.804

**다중 라운드 대화 규모.** 다양한 데이터 규모의 영향을 연구하기 위해 *SocraTeach*에서 25%, 50%, 75%의 다중 라운드 대화를 무작위로 추출하고, DTS 파이프라인을 여러 번 실행하여 125% 규모의 대화로 확장한 후 *Socrat-icLM*을 훈련시켰습니다. 그림 4의 결과는 (i) 본 데이터가 다양한 규모에서 효과적일 뿐만 아니라 영향력도 있어 LLM의 교육 능력을 크게 향상시킬 수 있음을 나타냅니다. (ii) 데이터 양이 증가함에 따라 교육 능력도 비례하여 증가하는 것을 관찰할 수 있습니다. 이러한 상관관계는

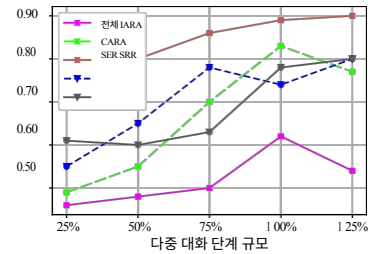


그림 4: 대화 규모 효과.

모델 성능에서 데이터 양의 중요성. 특히, 최소 75%( $\approx 26K$ )의 대화 규모가

GPT4의 *종합 품질*을 뛰어넘기 위해서는 로그 데이터가 필요합니다. (iii) 데이터 양이 35K 임계값을 초과하면 포화점에 접근하는 경향이 있으며, 이 시점 이후 데이터 양을 더 늘려도 모델 성능 향상에 미치는 증분 효과는 작아집니다. 구체적으로, 125% 데이터 규모에서 *LARA* 지표는 하락세를 보이며, 이러한 포화 상태의 근본 원인이 모델의 잘못된 답변 식별 능력 감소(*전체 품질* 하락은 그에 따른 결과임)을 시사합니다. 이는 다중 라운드 대화 데이터가 증가함에 따라 "*잘못된 답변*"에 대한 단일 라운드 대화 데이터의 비율이 감소하기 때문일 수 있습니다. 다중 대화 데이터 규모가 125%를 초과할 때 이 비율이 특정 임계값 아래로 떨어지면 효과성이 감소할 수 있습니다.

**문제 해결 데이터의 규모.** 그림 5는 문제 해결 데이터와 대화 데이터 간의 비율  $\alpha$ 를 조정함에 따른 *SocraticLM*의 성능 변화를 보여줍니다. 이 추세는 문제 해결 데이터가 너무 적거나 너무 많을 경우 만족스러운 문제 해결 능력을 이끌어내지 못함을 시사합니다.

문제 해결 데이터와 대화 데이터 간의 비율  $\alpha$ 를 조정함에 따라 *SocraticLM*의 성능 변화가 *나타납니다*. 이 추세는 문제 해결 데이터가 너무 적거나 너무 많을 경우 만족스러운 문제 해결 능력을 이끌어내지 못함을 시사합니다. 대신 교육 대화 데이터와의 균형을 맞출 필요가 있습니다. 실제로 문제 해결 데이터를 과도하게 도입하면 GSM8K 정확도가 1.9% 하락할 수 있다. 이는 *SocraticLM*의 문제 해결 능력에 해당하는 파라미터가 교육 대화로 초기 미세 조정된 후 교란을 받을 수 있기 때문으로 해석된다. 재훈련 시

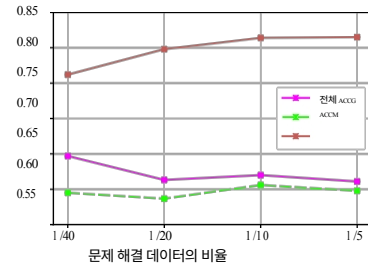


그림 5: 문제 해결 데이터의 효과  
데이터 규모에 따른 문제 해결 효과.

문제 해결 데이터가 있는 경우, 이 능력의 과소적합과 과적합 사이의 미묘한 균형을 재조정해야 합니다.

## 7 결론

본 논문에서는 소크라테스식 "사고를 자극하는" 맞춤형 교육을 촉진하기 위해 설계된 대규모 언어 모델(LLM)인 *SocraticLM*을 소개한다. *SocraticLM* 구축을 위해 우리는 "학생-교사-학생" 파이프라인을 제안하여 *SocraTeach* 데이터셋을 구축했는데, 이는 여섯 가지 학생 인지 상태를 시뮬레이션하고 네 가지 핵심 교수 능력을 강화했다. 또한 LLM을 위한 포괄적인 교수 능력 평가 시스템을 개발하였다. 실험 결과 *SocraticLM*은 GPT-4 등 기존 LLM을 현저히 능가하는 성능을 보였으며, *SocraTeach* 데이터셋 내 각 구성 요소의 필요성을 검증하였다. 부록 H, I, J에서는 추가 사례, 광범위한 영향, 한계점 및 향후 연구 방향을 논의한다.

## 감사와 자금 지원 공개

본 연구는 중국 국가 핵심 연구 개발 프로그램(No.2021YFF0901005), 중국 국가 자연 과학 기금(No.62477044, 62337001), 안후이성 핵심 기술 연구 개발 프로그램(No.202423k09020039)의 지원으로 부분적으로 수행되었습니다..

## 참고문헌

- [1] 조쉬 아치암, 스티븐 애들러, 산디니 아가르왈, 라마 아마드, 일게 아카야, 플로렌시아 레오니 알레만, 디오고 알메이다, 안코 알텐슈미트, 샘 알트만, 시암 알 아나드카트 외. GPT-4 기술 보고서. *arXiv 사전 인쇄본 arXiv:2303.08774*, 2023.
- [2] 투판 아디구젤, 메흐메트 할둔 카야, 파티흐 큐르산 칸수. 인공지능으로 교육 혁명 일으키기: ChatGPT의 변혁적 잠재력 탐구. *현대 교육 기술*, 2023.

- [3] Geoffrey D Borman, Gina M Hewes, Laura T Overman, and Shelly Brown. 포괄적 학교 개혁과 학업 성취도: 메타분석. *교육 연구 리뷰*, 73(2):125–230, 2003.
- [4] Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Oeistein Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant 외. 언어 교육 및 평가 기술에 대한 대규모 언어 모델의 적용에 관하여. *arXiv 사전 인쇄본* *arXiv:2307.08393*, 2023.
- [5] 이은 첸, 레이 황, 한신 첸, 위엔시엔 첸, 리앙이 리. Gptutor: 코드 설명을 위한 ChatGPT 기반 프로그래밍 도구. *국제 인공지능 교육 학회*, 321–327쪽. Springer, 2023.
- [6] W ei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez 외. Vicuna: ChatGPT 품질의 90%\*를 구현한 오픈소스 채팅봇으로 GPT-4를 압도하다. *https://vicuna.lmsys.org (2023년 4월 14일 접속)*, 2(3):6, 2023.
- [7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma 외. 명령어 미세 조정 언어 모델의 확장. *arXiv 사전 인쇄본* *arXiv:2210.11416*, 2022.
- [8] 칼 코베, 비닛 코사라주, 모하마드 바바리안, 마크 천, 준희우, 루카시 카이저, 마티아스 플라퍼트, 제리 트워렉, 제이콥 힐튼, 나카모 레이이치로 외. 수학 단어 문제 해결을 위한 검증기 훈련. *arXiv 사전 인쇄본* *arXiv:2110.14168*, 2021.
- [9] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, Guanliang Chen. 대규모 언어 모델이 학생들에게 피드백을 제공할 수 있을까? ChatGPT 사례 연구. *2023 IEEE 국제 첨단 학습 기술 컨퍼런스(ICALT)*, 323–325쪽. IEEE, 2023.
- [10] 단유하오, 레이즈카이, 구이양, 리용, 인장하오, 린자주, 예린하오, 티에즈이안, 저우유젠, 왕이레이 외. 에듀챗: 지능형 교육을 위한 대규모 언어 모델 기반 챗봇 시스템. *arXiv 사전 인쇄본* *arXiv:2308.02773*, 2023.
- [11] 이안 제이 디어리, 스티브 스트랜드, 폴린 스미스, 크레스 페르난데스. 지능과 교육 성취도. *Intelligence*, 35(1):13–21, 2007.
- [12] 정샤오 두, 위지에 치엔, 샤오 리우, 밍 딩, 지예중 치우, 지린 양, 지예 탕. GLM: 자동회귀 공백 채우기를 통한 일반 언어 모델 사전 훈련. *제60회 전산언어학회 연차대회 논문집 (제1 권: 장문 논문)*, 320–335쪽, 2022.
- [13] Linda Elder, Richard Paul. 사고, 교수, 학습에서 소크라테스식 질문의 역할. *The Clearing House*, 71(5):297–301, 1998.
- [14] 에이미 S. 핀, 매튜 A. 크래프트, 마틴 R. 웨스트, 줄리아 A. 레너드, 크리스탈 E. 비시, 레베카 E. 마틴, 마거릿 A. 셰리던, 크리스토퍼 F.O. 가브리엘리, 존 D.E. 가브리엘리. 인지 능력, 학생 성취도 평가, 그리고 학교. *심리과학*, 25(3):736–744, 2014.
- [15] 마렌 포르마진, 올리히 슈뢰더스, 올라프 쾰러, 올리버 빌헬름, 한스 베스트마이어. 심리학 전공 학생 선발. *심리학 리뷰*, 2011.
- [16] Wensheng Gan, Zhenlian Qi, Jiayang Wu, Jerry Chun-Wei Lin. 교육 분야의 대규모 언어 모델: 비전과 기회. *2023 IEEE 국제 빅데이터 컨퍼런스(BigData)*, 4776–4785쪽. IEEE, 2023.
- [17] Arthur C Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, Max M Louwerse. Autotutor: 자연어 대화 기능을 갖춘 튜터. *행동 연구 방법, 도구 및 컴퓨터*, 36:180–192, 2004.
- [18] 한지은, 유하늘, 김윤수, 명준호, 김민선, 임현승, 김주호, 이탁연, 홍화정, 안소연 외. 레시퍼: EFL 쓰기 교육에 ChatGPT를 통합하는 방법. *제10회 ACM 대규모 학습 컨퍼런스 논문집*, 416–420쪽, 2023.

- [19] 에릭 A. 하누셰크, 스티븐 G. 리브킨. 교사 역량 가치 부가 측정 활용에 관한 일반화. *American economic review*, 100(2):267–271, 2010.
- [20] 크리스토테아 헤로도투, 바트 리엔티스, 아비나쉬 보로와, 즈데넥 즈드라할, 마틴 홀스타. 고등교육에서의 예측 학습 분석 대규모 구현: 교사의 역할과 관점. *교육 기술 연구 및 개발*, 67:1273–1306, 2019.
- [21] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen 외. Lora: 대규모 언어 모델의 저순위 적응. *국제 학습 표현 학회*, 2021.
- [22] Liya Hu, Zhiang Dong, Jingyuan Chen, Guifeng Wang, Zhihua Wang, Zhou Zhao, Fei Wu. Ptadisc: 콜드스타트 시나리오에서 맞춤형 학습을 지원하는 교차과목 데이터셋. *신경정보처리시스템 발전*, 36:44976–44996, 2023.
- [23] 황전야, 류치, 천위잉, 우러, 샤오켈리, 천엔홍, 마하이핑, 후구오핑. 학습인가 망각인가? 학생의 지식 숙련도 추적을 위한 동적 접근법. *ACM Transactions on Information Systems (TOIS)*, 38(2):1–33, 2020.
- [24] 진 제니퍼, 김 미라. 프로그래밍 언어 학습을 위한 GPT 기반 맞춤형 이러닝 시스템. *응용과학*, 13(23):12773, 2023.
- [25] 타니샤 조시, 제시카 스토크스-패리시, 레이첼 싱글턴, 마이클 토도로비치. 생성형 인공지능 대규모 언어 모델로 강화된 의학 교육. *Trends in Molecular Medicine*, 2023.
- [26] 엔켈레이다 카스네치, 카트린 제슬러, 슈테판 퀴헤만, 마리아 바네르트, 다리나 데멘티예바, 프랑크 피셔, 우르스 가셔, 게오르크 그로, 슈테판 권네만, 아이케 헬러마이어 외. ChatGPT를 선한 목적으로? 교육 분야 대규모 언어 모델의 기회와 도전. *학습과 개인차*, 103:102274, 2023.
- [27] 릭 콘첼-케지오르스키, 수브로 로이, 아이다 아미니, 네이트 쿠시먼, 한나네 하지시르지. Mawps: 수학 단어 문제 저장소. 2016년 북미계산언어학회(ACL) 인간 언어 기술 컨퍼런스 논문집, 1152–1157쪽, 2016.
- [28] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, Bernard Ghanem. Camel: 대규모 언어 모델 사회의 "마음" 탐색을 위한 의사소통 에이전트. *신경정보처리시스템 발전*, 36, 2024.
- [29] 리칭야오, 푸링웨, 장웨이밍, 천센위, 위징웨이, 샤웨이, 장웨이난, 탕루이밍, 위웅. 교육용 대규모 언어 모델 적용: 기초 역량, 잠재력 및 과제. *arXiv 사전 인쇄본 arXiv:2401.08664*, 2023.
- [30] 리즈종, 데릭 호이엔. 잊지 않고 학습하기. *IEEE 패턴 분석 및 기계 지능 거래지*, 40(12):2935–2947, 2017.
- [31] 진위 린. Rouge: 요약문 자동 평가 패키지. *텍스트 요약의 분화*, 74–81쪽, 2004.
- [32] 신 린, 진야 황, 흥커 자오, 엔홍 천, 치 류, 데푸 렌, 신 리, 하오 왕. 수학 단어 문제 해결을 위한 관계 강화 계층적 솔버 학습. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [33] 류자위, 왕페이, 마하이핑, 황젠야, 류치, 천엔홍, 수유. 온라인 학습 시스템에서의 시간적 인지 진단을 위한 확률적 프레임워크. *컴퓨터 과학 및 기술 저널*, 38(6):1203–1222, 2023.
- [34] 류지야유, 황젠야, 마즈위안, 류치, 천홍, 수토허, 류하이핑. 상식적 공식 지식 습득을 통한 수학적 추론 유도. 제29회 ACM SIGKDD 지식 발견 및 데이터 마이닝 컨퍼런스 논문집, 1477–1488쪽, 2023.

- [35] 류치, 황젠야, 인위, 천옌훙, 송후이, 수위, 후귀핑. Ekt: 학생 성과 예측을 위한 연습 인식 지식 추적. *IEEE Transactions on Knowledge and Data Engineering*, 33(1):100–115, 2019.
- [36] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: 프롬프트 튜닝은 규모와 작업 전반에 걸쳐 파인 튜닝과 비교 가능한 성능을 가질 수 있다. *제60회 전산언어학회 연차대회 논문집 (제2권: 단편 논문)*. 전산언어학회, 2022.
- [37] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei 외. FLAN 컬렉션: 효과적인 지시 튜닝을 위한 데이터 및 방법 설계. *국제 기계 학습 컨퍼런스*, 22631–22648쪽. PMLR, 2023.
- [38] 윤 뢰, 진 양, 판동 명, 야푸 리, 지에 저우, 위얼 장. 지속적 미세 조정 중 대규모 언어 모델의 재앙적 망각에 관한 실증 연구. *arXiv 사전 인쇄본 arXiv:2308.08747*, 2023.
- [39] 마하이핑, 왕창첸, 주형슈, 양상상, 장샤오밍, 장싱이. 상호작용하지 않은 연습을 통한 인지 진단 향상: 협업 인식 혼합 샘플링 접근법. *AAAI 인공지능 컨퍼런스 논문집*, 제38권, 8877–8885쪽, 2024.
- [40] Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, Mrinmaya Sachan. Mathdial: 수학적 추론 문제에 기반한 풍부한 교육학적 특성을 지닌 대화형 튜터링 데이터셋. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, 2023.
- [41] Julia M Markel, Steven G Opferman, James A Landay, Chris Piech. Gpteach: GPT 기반 학생들과의 대화형 TA 훈련. *제10회 ACM 대규모 학습 컨퍼런스 논문집*, 226–236쪽, 2023.
- [42] 스티븐 무어, 리처드 통, 안잘리 싱, 류 지타오, 후 시양젠, 루 위, 리양 줄린, 차오 천, 하산 코스라비, 폴 데니 외. 대규모 언어 모델로 교육 역량 강화: 차세대 인터페이스 및 콘텐츠 생성. *《인공지능 교육 국제 학회》*, 32–37쪽. 스프링거, 2023.
- [43] 키쇼어 파피네니, 살림 루코스, 토드 워드, 위징 주. Bleu: 기계 번역 자동 평가 방법. *제40회 전산언어학회 연차대회 논문집*, 311–318쪽, 2002.
- [44] 박민주, 김소정, 이승현, 권순우, 김규석. 학생 모델링을 활용한 대화 기반 튜터링 시스템을 통한 맞춤형 학습 강화. *arXiv 사전 인쇄본 arXiv:2403.14071*, 2024.
- [45] Richard Paul, Linda Elder. 비판적 사고: 소크라테스식 질문의 기술. *발달 교육 저널*, 31(1):36, 2007.
- [46] 루안 세리, 장 리웨이, 쉬 저스틴, 탐 브라이스 조쿤, 추 정녕, 주 예상, 머넌 엘리자베스 L., 브런스킬 엠마, 랜데이 제임스 A. 퀴즈봇: 사실적 지식을 위한 대화 기반 적응형 학습 시스템. *2019 CHI 인간-컴퓨터 상호작용 학회 논문집*, 1–13쪽, 2019.
- [47] 시위취와 취샤오웨이. 인지 능력이 학업 성취도에 미치는 영향 분석: 자기 모니터링의 조절 역할. *프론티어스 인 사이콜로지*, 13:996504, 2022.
- [48] 아로히 스리바스타바, 데니스 클레이조, 우즈이. 모방 게임을 넘어: 언어 모델의 능력 정량화 및 외삽. *기계 학습 연구 논문집*, (5), 2023.
- [49] Katherine Stasaski, Kimberly Kao, Marti A Hearst. Cima: 교육용 튜터링을 위한 대규모 오픈 액세스 대화 데이터셋. *제15회 교육 애플리케이션 구축을 위한 NLP 혁신적 활용 워크숍 논문집*, 52–64쪽, 2020.

- [50] 쉰젠원, 저우루이, 리앙루샤, 가오루, 리우산뉴야, 리칭, 장카이, 장루루. 양상블 지식 추적: 학습 과정에서의 상호작용 모델링. *Expert Systems with Applications*, 207:117680, 2022.
- [51] Anais Tack, Chris Piech. AI 교사 테스트: 교육 대화에서 Blender와 GPT-3의 교육적 능력 측정. *국제 교육 데이터 마이닝 학회*, 2022.
- [52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale 외. Llama 2: 오픈 파운데이션 및 미세 조정된 채팅 모델. *arXiv 사전 인쇄본 arXiv:2307.09288*, 2023.
- [53] 왕페이, 류치, 천엔홍, 황젠야, 인위, 왕시진, 수위. NeuralCD: 인지 진단을 위한 일반 프레임워크. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8312–8327, 2022.
- [54] 왕 롱위, 류 천양, 지 텐보, 장 지루이, 위 디안, 시 슈밍, 투 자오펑. 대규모 언어 모델을 활용한 문서 수준 기계 번역. *2023년 자연어 처리 경험적 방법론 컨퍼런스*, 2023.
- [55] 제이슨 웨이, 마르텐 보스마, 빈센트 자오, 켈빈 구, 애덤스 웨이 유, 브라이언 레스터, 난 두, 앤드류 M. 다이, 관 V. 레. 미세 조정된 언어 모델은 제로샷 학습자이다. *국제 표현 학습 컨퍼런스*, 2021.
- [56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou 외. 사고의 사슬 프롬프팅은 대규모 언어 모델에서 추론을 유도한다. *신경 정보 처리 시스템 발전*, 35:24824–24837, 2022.
- [57] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 사고의 나무: 대규모 언어 모델을 활용한 의도적 문제 해결. *신경정보처리시스템 발전*, 36, 2024.
- [58] 장량, 린종하오, 광즈이, 쉬성, 모하메드 예신, 후시안겐. SPL: 대규모 언어 모델 기반 학습을 위한 소크라테스식 학습 플랫폼. *arXiv 사전 인쇄본 arXiv:2406.13919*, 2024.
- [59] 웨인 신 자오, 쿤 저우, 준이 리, 티안이 탕, 샤오레이 왕, 유펑 후, 잉치안 민, 베이천 장, 준지에 장, 지첸 동 외. 대규모 언어 모델에 대한 조사. *arXiv 사전 인쇄본 arXiv:2303.18223*, 2023.

## A 문제 분해에 관한 추가 세부 사항

GSM8K 원시 데이터에는 각 문제에 대한 분해가 제공되었으며, 이는 우리의 요구 사항을 대체로 충족하는 것으로 확인되었습니다. 따라서 우리는 A.1의 프롬프트를 사용하여 GPT4로 MAWPS의 문제를 분해합니다.

### A.1: 문제 분해 프롬프트

당신은 수학 교사이며, 제가 제시하는 수학 문제를 단계별로 해결해 주시기 바랍니다. 각 단계에서 먼저 안내 질문을 제시한 후, 다음 기준을 충족하며 해결 과정을 설명해야 합니다:

- 모든 단계를 거친 후 최종 결과는 정답과 동일해야 합니다
- 단계는 가능한 한 최소화해야 합니다
- 계산 과정 자체를 단계로 제시하지 마십시오
- 단계로 결과 요약이 아니어야 합니다
- 각 단계는 '\n'으로 구분됩니다
- 각 단계가 잘 정리되고 정확하며 따라하기 쉬워야 하며, 각 단계는 이전 단계를 기반으로 해야 합니다.

#문제: 나탈리아는 4월에 친구 48명에게 클립을 팔았고, 5월에는 그 절반만큼 팔았습니다. 나탈리아가 4월과 5월에 총 몇 개의 클립을 팔았나요?

#답변: 72

#단계: 나탈리아가 5월에 판매한 클립 수는? \*\* 나탈리아는 5월에  $48/2 = 24$ 개의 클립을 판매했습니다.\n나탈리아가 4월과 5월에 총 몇 개의 클립을 판매했나요? \*\* 나탈리아는 4월과 5월에 총  $48+24 = 72$ 개의 클립을 판매했습니다.

#문제: {여기 목표 문제가 있습니다.} #답변: {여기 목표 답변이 있습니다.}

## B "학장-교사-학생" 파이프라인 프롬프트

### B.1: 학생에이전트 프롬프트

당신은 초등학생입니다. 대화를 참고하여 가능한 답변을 생성해 주세요. 답변을 생성할 때는 먼저 다음 상황 중 하나를 선택한 후, 해당 상황에 맞는 학생의 입장에서 다음 기준을 충족하도록 답변해야 합니다:

- 선생님이 아니므로, 교사 답변을 하지 마세요!
- 질문 속 인물에 대한 질문을 하지 마십시오.
- 답변은 주어진 대화 내용과 일관성이 있어야 합니다.

상황: (1) 학생이 문제의 의미를 이해하지 못하는 경우; (2) 학생이 교사가 설명한 내용을 이해하지 못하는 경우; (3) 학생이 계산에서 실수를 하는 경우; (4) 학생의 지식 습득이 부족한 경우; (5) 학생의 학습 의욕이 약한 경우; (6) 학생이 모든 면에서 능력이 뛰어난 경우.

다음은 몇 가지 예시입니다:

#문제: 나탈리아는 4월에 친구 48명에게 클립을 팔았고, 5월에는 그 절반만큼 팔았습니다. 나탈리아가 4월과 5월에 모두 몇 개의 클립을 팔았나요?

#대화: [교사]나탈리아가 5월에 클립을 몇 개 팔았나요?[학생]수학이 잘 안 돼요. 다시 설명해 주실 수 있나요?[교사]전혀 문제없어요. 단계별로 풀어보자. 나탈리아가 5월에 4월 판매량의 절반만큼 클립을 팔았다면, 4월에 48개를 팔았으니 5월에는 몇 개를 팔았을까? 4월 판매량에서 절반을 구해보자. 몇 개가 나올까? #응답:상황 (3) 선택[학생]알겠어요, 48의 절반은 240이에요.

...

목표 문제는 다음과 같습니다:

#문제: {여기 목표 문제가 있습니다.} #대화: {여기 목표 대화가 있습니다.}

## B.2: 교사에이전트의 프롬프트

저는 초등학생입니다. 선생님은 항상 소크라테스식 질문법으로 답변하시는 분이십니다. 정답을 직접 알려주지 않으시고, 제가 스스로 생각할 수 있도록 항상 적절한 질문만 던지려 하십니다. 문제를 제 수준에 맞도록 더 단순한 부분으로 나누어 설명해 주셔야 합니다. 위의 단계를 [단계 1]부터 질문 형식으로 설명해 주세요. 설명을 마쳤다고 생각되면 답변 끝에 "[END]"를 적어 주세요.

### (학생의 상황에 따라 다음 예시 중 하나를 선택): (상황 1)

#문제: 나탈리아는 4월에 친구 48명에게 클립을 팔았고, 5월에는 그 절반만큼 팔았습니다. 나탈리아가 4월과 5월에 총 몇 개의 클립을 팔았나요?

#단계: [단계 1] 나탈리아가 5월에 클립을 몇 개 팔았나요? [단계 2] 나탈리아가

나탈리아가 4월과 5월에 모두 몇 개의 클립을 팔았나요?

#대화:[교사]: 나탈리아가 5월에 클립을 몇 개 팔았나요? [학생]: 혼란스러워요. "반만큼"이 4월과 같은 수라는 뜻 인가요?

#응답:[교사]: 사실 "반만큼"은 그 수의 절반을 의미합니다. 그러면 나탈리아가 4월에 클립 48개를 팔았다면, 5월에는 몇 개를 팔았을까요?

### (상황 2)

#문제: 랜디의 농장에는 망고나무가 60그루 있습니다. 코코넛나무는 망고나무의 절반보다 5그루 적습니다. 랜디의 농장에 총 몇 그루의 나무가 있을까요?

#단계: [단계 1] 랜디가 가진 망고나무는 몇 그루인가요? [단계 2] 랜디가 가진 코코넛나무는 몇 그루인가요? [단계 3]

랜디의 농장에 총 몇 그루의 나무가 있나요? #대화:[교사]: 랜디가 가진 망고나무는 몇 그루인가요? [학생]: 총 몇 그루인

지, 아니면 망고나무만 말하는 건가요?

#응답:[교사]: 지금은 망고나무만 이야기하고 있어요. 그러니까 망고나무만 고려해서, 문제에 따라 랜디가 몇 그루 가지고 있는지 말해줄 수 있나요?

### (상황 3-6)

...

다음은 대상 문제입니다:

#문제: {여기 목표 문제가 있습니다.} #단계: {여기 목

표 단계가 있습니다.} #대화: {여기 목표 대화가 있습

니다.}

## B.3: E/에이전트 프롬프트

당신은 교육 감독관입니다. 다음은 소크라테스식 교수법의 교육 과정입니다. 마지막 라운드에서 [교사]의 응답을 평가하 고 다음 교육 기준을 충족하는지 판단하십시오:

- [학생]이 상식적인 질문을 하거나 개념 이해에 어려움을 겪을 경우, [교사]는 직접 설명을 제공할 수 있습 니다.
- 그 외의 경우, 문제에 대한 직접적인 답변이나 해결책을 선언적 또는 수사적 방식으로 제시해서는 안 됩니다.
- [학생]이 실수를 했을 경우, 질문을 통해 어떤 오류를 범했는지 암시해야 합니다.
- [교사]는 교사답게 말해야 하며 "내가 물어볼게" 같은 표현을 사용하지 않습니다.
- 응답은 매끄럽고 유창해야 합니다.

응답이 기준을 충족한다고 생각되면 [True]로 답하고, 그렇지 않으면 [False]로 답하며 판단 근거를 설명하고 기준에 따라 수정하십시오.



**예시:**

#문제: 나탈리아는 4월에 친구 48명에게 클립을 팔았고, 5월에는 그 절반만큼 팔았습니다. 나탈리아가 4월과 5월에 총 몇 개의 클립을 팔았나요?

#대화:[교사]: 나탈리아가 5월에 클립을 몇 개 팔았나요?  
[학생]: 혼란스러워요. "반만큼"이 4월과 같은 수를 의미하나요?  
[교사]: 정확히는 아니에요. "반만큼"은 절반의 양을 의미합니다. 나탈리아가 4월에 클립 48개를 팔았다면, 48의 절반은 얼마일까요? 그게 5월에 팔린 수를 알려줄 거예요. 풀어볼 수 있겠니?

#응답:[맞음].

#문제: 나탈리아는 4월에 친구 48명에게 클립을 팔았고, 5월에는 그 절반만큼 팔았습니다. 나탈리아가 4월과 5월에 총 몇 개의 클립을 팔았나요?

#대화:[교사]: 나탈리아가 5월에 클립을 몇 개 팔았을까?  
[학생]: 48에 48을 2로 나눈 걸 더하는 건가요? 그러면 72인가요?  
[교사]: 5월에 팔린 클립 수를 알아내기 위해 2로 나누는 건 맞아요. 그런데 이렇게 물어볼게: 48에 48을 2로 나눈 값을 더하는 건가, 아니면 먼저 48 자체를 2로 나누는 건가?

#응답:[틀림]. [교사]의 답변은 부정확하며 [학생]의 계산 오류를 지적하지 못했습니다. [수정된 교사]: 방향은 맞았어! 하지만 계산 실수를 할 수 있거든. 48 더하기 48을 2로 나눈 결과가 얼마인지 다시 계산해 볼 수 있겠니?

#문제: 베티는 60세이며 가족 중 가장 나이 많은 사람입니다. 그녀의 딸은 그녀보다 40% 젊고, 손녀는 어머니 나이보다 1/3 더 어립니다. 손녀의 나이는 몇 살일까요?

#대화:[교사]: 베티의 딸은 몇 살인가요?  
[학생]: 60의 60%는... 36이니, 36인가요?  
[교사]: 백분율을 생각한 건 맞지만, 누군가가 "40% 더 어리다"고 할 때는 그 백분율을 100%에서 빼야 한다는 걸 기억하세요. 이 점을 염두에 두고 다시 계산해 볼 수 있겠니?

#응답:[틀림]. [학생]은 "베티의 딸 나이는?"에 대한 답을 정확히 구했지만, [교사]는 여전히 같은 질문을 합니다. [수정된 교사]: 네! 베티 딸의 나이를 정확히 구했어요. 이제 베티의 손녀 나이를 계산해 볼 수 있겠네요. 한번 해볼래요?

**목표 문제는 다음과 같습니다:**

#문제: {여기 목표 문제가 있습니다.} #대화: {여기 목표 대화가 있습니다.}

## C 교수 능력 향상에 관한 추가 정보

"무관한" 학생 응답을 위한 단일 라운드 대화를 구축하기 위해, 먼저 DTS 파이프라인으로 구성된  $Dia_M$ 에서 2,000개의 대화를 무작위로 선택합니다. 각 대화에서 학생 응답의 한 라운드를 무작위로 선택하여 MOOC에서 수집한 200개의 질문 중 무작위로 선택한 질문으로 대체합니다. 그런 다음 교장감독 하에 교사 에이전트가 질문에 답변을 거부하도록 하여, 최종적으로 2,000개의 단일 라운드 학생-교사 대화를 형성합니다.

### C.1: 질문 프롬프트

#문제: {여기 대상 문제가 있습니다.} #답변: {여기 대상 답변이 있습니다.} #분석: {여기 대상 분석이 있습니다.}  
#대화 기록: {여기 대상 대화 기록이 있습니다.}

당신은 초등학생입니다. 마지막 [대화 기록] 라운드에 대해 초등학생이 간결한 언어로 물어볼 수 있는 세 가지 질문을 제시하세요. 서로 다른 질문은 \*\*로 구분해야 합니다.

"잘못된 답에 의존"을 위한 다섯 가지 오답을 구성하기 위해, 한편으로는 학생의 답변에서 모든 숫자(예: "2")와 연산자(예: "+")를 식별하고, 그 안에 무작위로 변동을 도입합니다.

## C.2: (정확/오류) 응답프롬프트

당신은 초등학생입니다. [대화 기록]의 마지막 라운드 [학생] 답변을 다음 기준에 맞춰 (5개의 오답 포함) 2개의 정답으로 재구성해 주세요:

- 지난 라운드에서 [선생님]이 제기한 질문에만 답하고, 다른 것은 묻지 마십시오.
- 초등학생의 어조로 답변하십시오.
- 다른 답변은 \*\*로 구분하십시오.

#문제: {여기 목표 문제가 있습니다.} #답변: {여기 목표 답변  
이 있습니다.} #분석: {여기 목표 분석이 있습니다.}

#대화 기록: {대상 대화 기록입니다.}

문제 수	11,147
다중 라운드 대화 수 <sub>M</sub>	35,151
단일 라운드 대화 수( <i>Dia<sub>S</sub></i> )	22,000
단일 라운드 대화 총 개수	207,581
최대 / 최소 라운드 수	12 / 3
단계별 안내 질문 평균 개수	3.29
평균라운드 수 <sub>M</sub>	5.28
학생 응답 평균 길이	16.4
교사 응답 평균 길이	30.3

표 3: SocraTeach 데이터셋 통계.

10에서 한 숫자로 범위를 조정하거나 임의로 연산자를 다른 연산자로 대체합니다. 이러한 규칙을 적용하여 두 개의 새로운 학생 응답을 생성합니다. 한편, GPT4를 사용하여 원본 학생 응답을 재작성하고 C.2의 프롬프트를 활용해 세 개의 오답을 생성합니다. 이 다섯 개의 응답을 바탕으로 교사 응답을 유도하여 10,000건의 학생-교사 대화를 확보합니다.

## D SocraTeach 데이터셋 통계

제3.5절에서 언급한 바와 같이, 본 연구의 SocraTeach 데이터셋은 35,000건의 다중 라운드 대화(*Dia<sub>M</sub>*)와 22,000건의 단일 라운드 대화(*Dia<sub>S</sub>*)로 구성됩니다. *Dia<sub>M</sub>*의 평균 대화 라운드 수는 5.28회입니다. 학생/교사의 응답은 평균적으로 각각 16.4단어/30.3단어를 포함합니다. 전체 통계는 표 3에 요약되어 있습니다. 또한 그림 6(a)에서 볼 수 있듯이, 대부분의 교사 대화는 5~6라운드로 구성됩니다. 그림 6(b)에서는 SocraTeach 내 학생 인지 상태의 확률 질량 함수를 시각화하였다. 여기서 (1)-(5)는 제한된 학생 인지 시스템(3.3절)의 차원 중 하나에서 취약한 학생 프로파일을 나타내며, (6)은 모든 차원에서 우수한 상태를 가진 학생을 나타낸다. 학생 에이전트는 모든 측면에서 우수한 학생을 시뮬레이션하거나, 계산 능력(즉, "(3)") 또는 지식 숙달도(즉, "(4)")에 문제가 있는 학생을 시뮬레이션하는 경향이 있음을 알 수 있습니다.

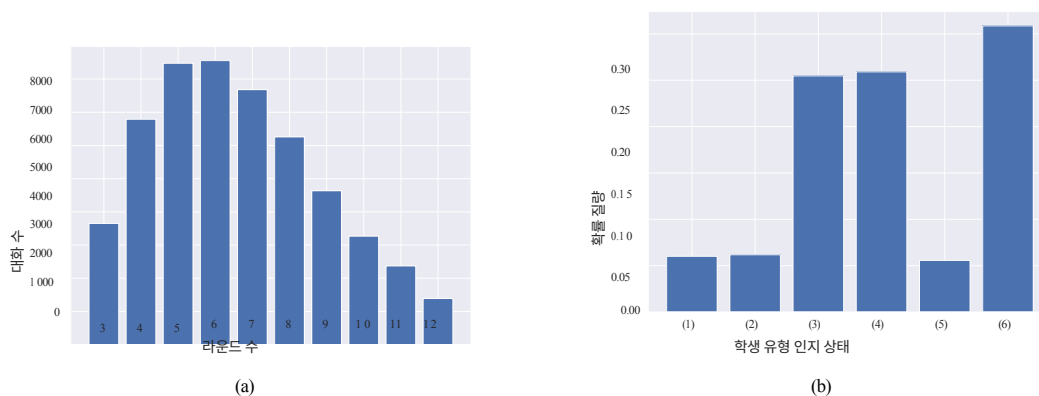


그림 6: 라운드 수(a) 및 학생 인지 상태(b)의 분포.

## E 지시 조정 템플릿

그림 7은 대화 데이터와 문제 해결 데이터를 가르치기 위한 지시 템플릿을 보여줍니다.

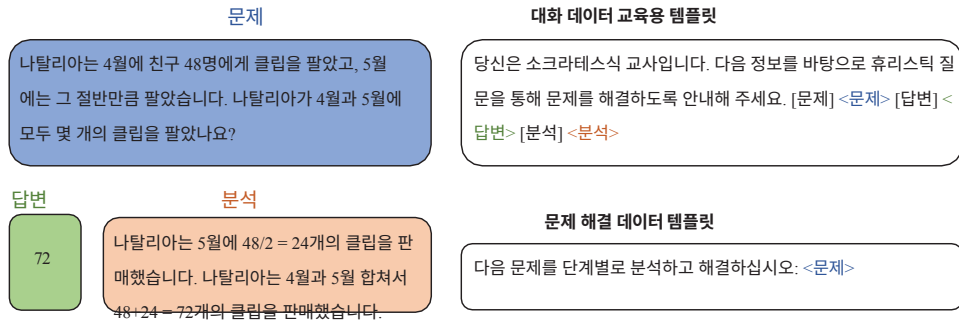


그림 7: 지시문 조정용 템플릿.

## F 교육 품질에 대한 인간 평가

다양한 대규모 언어 모델(LLM)의 *전반적인 품질*을 평가하기 위해, 우리는 10명의 고학력 어노테이터를 모집하여 각 모델이 제공한 지시문을 GPT-4와 비교하여 순위를 매기도록 했습니다. 그림 8은 해당 템플릿을 보여줍니다.

대화 ID:

교육 품질 평가

1. 교사가 학생에게 수학 문제를 설명하는 다음 대화를 읽어보십시오.

2. 학생의 마지막 응답에 대해, 네 가지 측면에서 어느 교사의 지도가 더 나은지 평가하십시오:

- 이해력: 교사가 학생의 의도를 정확히 이해했는지 평가하십시오.
- 설명: 교사가 학생의 문제를 효과적으로 해결하고 적절하며 실행 가능한 지도를 제공했는지 여부.
- 언어: 교사의 지도가 소크라테스식 교수법의 요구사항(질문 형식으로 제시되고 직접적인 답을 주지 않음)을 따르는지 평가하십시오.
- 가독성: 교사의 지시가 읽기 쉽고 지나치게 직설적이지 않은지 평가하십시오.

#문제: 한 가게에서 우노 자이언트 패밀리 카드 한 세트 가격이 12달러입니다. 이반이 열 세트를 샀을 때, 각 세트당 2달러 할인을 받았 습니다. 이반이 총 얼마를 지불했나요?

#답변: 100달러.

#분석: 우노 자이언트 패밀리 카드 한 장당 가격은 12달러가 아닌 12달러 - 2달러 = 10달러입니다. 이반은 우노 자이언트 패밀리 카드 10장을 10달러 x 10 = 100달러에 구입했습니다.

#단계: (1) 우노 자이언트 패밀리 카드 한 장의 가격은 얼마인가요?  
(2) 이반이 총 얼마를 지불했나요?

#대화 기록:  
[교사]: 우노 자이언트 패밀리 카드 한 장은 얼마인가요? [학생]: 왜 할인을 계산해야 하나요?  
[교사]: 좋은 질문이에요! 할인을 받고 물건을 사면 원래 가격보다 적게 내게 되거든요. 그래서 우노 자이언트 패밀리 카드의 원래 가격 을 알면 이반이 할인으로 얼마나 절약했는지 이해하는 데 도움이 돼요. 이제 이걸 알고, 할인을 적용한 후 이반이 총 얼마를 지불했는지 알아볼 수 있겠어요?  
[학생]: 이해가 안 돼요. 할인이 카드 한 장당 가격을 어떻게 바꾸나요?  
  
[교사 A]:  
할인은 카드 원래 가격에서 깎아주는 것과 같아요. 이 경우, 카드 한 장당 원래 가격이 12달러고 2달러 할인이 적용된다면, 원래 가격에 서 할인액을 빼야 해요. 그러면 카드 한 장당 새 가격은 12달러 - 2달러가 되겠죠. 그게 얼마일까요?  
  
[교사 B]:  
아주 좋은 질문이에요! '할인'이란 원래 가격에서 깎아주는 것을 의미합니다. 그렇다면 이반이 가진 10장의 우노 자이언트 패밀리 카드 각각에 2달러 할인을 받았다면, 카드 한 장당 얼마를 지불했을까요?

교사 A

동일

교사 B

선택

그림 8: *전반적 품질*에 대한 인간 평가 템플릿.

## G 구현 세부 사항

우리의 *SocraticLM*은 ChatGLM3-6b(ChatGLM3-6b-Base 아님)를 학습률 0.02, 배치 크기 64로 2 에포크 동안 P-Tuning[36]하여 얻었습니다. 데이터셋 분할은 5절에서 설명한 바와 같이, 객관적 교수 능력(*LARA*, *CARA*, *SER*, *SRR*) 평가용 단일 라운드 대화 400개와 주관적 *전반적 품질* 평가용 다중 라운드 대화 1,000개로 구성된 테스트 세트로 이루어집니다. *SocraTeach*의 나머지 데이터 중 10%/90%는 검증/훈련에 사용됩니다. 모든 실험은 NVIDIA RTX 3090 GPU 6개를 탑재한 서버에서 수행되었습니다. 데이터셋과 코드는 <https://github.com/Ljyustc/SocraticLM>에서 확인할 수 있습니다.

## H 추가 모델 샘플

다음은 학생의 "잘못된 답변", "올바른 답변", "질문형 답변", "관련 없는 답변"에 대한 *SocraticLM*과 GPT4의 네 가지 샘플입니다. 첫 번째 예시에서 *SocraticLM*과 GPT4 모두 학생의 오류를 식별하지만, GPT4는 "7"이라는 정답을 직접 제시하여 소크라테스식 교육의 요구사항과 상충됩니다. 두 번째 예시에서 GPT4는 거의 완성된 해결 과정을 제시하지만, 단계별 소크라테스식 사고 지도 요건을 충족하지 못합니다. 세 번째 예시에서 GPT4의 표현은 더 딱딱하며 실제 인간 교사의 느낌을 결여합니다. 네 번째 예시에서 *SocraticLM*은 학생이 먼저 수학 문제를 풀도록 유도하는 반면, GPT4는 관련 없는 질문에 답변하는 방향으로 전환됩니다.

### 샘플 1: 잘못된 답변

#문제: 브렌다의 어머니가 5명에게 줄 쿠키를 만들었습니다. 그녀가 35개의 쿠키를 준비했고 각 그룹이 같은 수의 쿠키를 받았다면, 각 그룹은 몇 개씩 받았을까요?

#대화 기록:

[교사]: 각 사람이 받는 쿠키 개수를 어떻게 계산할 수 있을까요? [학생]: 그냥 나눠 주시면 안 되나요?

[교사]: 음, 그렇게 할 수도 있지만, 왜 나누는지를 이해하는 게 중요해. 나누기는 특정 수의 사람이나 사물 사이에 무언가를 균등하게 분배할 수 있게 해줘. 이 경우, 브렌다 엄마가 5명을 위해 쿠키 35개를 만들었다면, 각 사람이 같은 양을 받도록 그 쿠키를 어떻게 나눠야 할까?

[학생]: 아, 그냥 쿠키 총 개수를 사람 수로 나누면 되는 건가요? [선생님]: 맞아! 이 방법으로 브렌다 엄마가 5명에게 35개의 쿠키를 만들었다면, 한 사람당 몇 개씩 받을 수 있을까?

[학생]: 잘 모르겠어요. 35를 5로 나누면 얼마죠?

[교사]: 좋은 질문이에요! 한번 생각해 볼까요? 5개 그룹이 있고, 35개를 똑같이 나눠야 한다면, 각 그룹에 몇 개씩 들어가야 할까요? 35에 5가 몇 번 들어가는지, 넘치지 않게 계산해 보는 걸로 시작해 볼 수 있어요. 한번 해볼래요?

[학생]: 그럼 한 사람당 쿠키 8개씩 받는다는 뜻인가요?

#### SocraticLM:

좋은 시작이네요! 하지만 다시 한번 확인해 볼까요. 35를 5로 나누면 무엇이 나오나요?

#### GPT4:

아직은 아닙니다. 35개의 물건으로 이루어진 한 무리를 상상해 보세요. 이를 5개의 동일한 무리로 나누면, 각 무리에는 실제로 7개의 물건이 포함된다는 것을 알 수 있습니다. 이 경우 각 사람은 7개의 쿠키를 받게 됩니다. 이해가 되시나요?

### 샘플 2: 정답

#문제: 제프는 12일 동안 매일 도넛 10개를 만듭니다. 제프는 매일 도넛 하나를 먹습니다. 그러자 크리스가 와서 도넛 8개를 먹었습니다. 도넛 10개가 한 상자에 들어간다면, 제프는 도넛으로 몇 개의 상자를 채울 수 있을까요?

#대화 기록:

[교사]: 제프가 도넛을 몇 개 만들까요?

[학생]: "만든다"에 그가 먹는 도넛도 포함되나요?

[교사]: "제프가 만든다"는 표현은 제프가 만든 도넛의 총량을 의미합니다. 그가 먹었는지 여부와 상관없이요. 그럼 제프가 총 몇 개의 도넛을 만들었나요? [학생]: 제프는 매일 10개의 도넛을 만들고 12일 동안 만들었으니, 총 120개의 도넛을 만들었습니다.

[교사]: 맞아요! 자, 그럼 제프가 먹는 도넛은 몇 개일까요?

[학생]: 제프는 매일 하나씩 먹어서 12일 동안 도넛을 만들었기 때문에 12개의 도넛을 먹습니다.

**SocraticLM:**

훌륭합니다! 자, 제프가 도넛 12개를 먹고 총 120개를 만들었다면, 그가 먹은 후 남은 도넛은 몇 개인가요?

남게 될까요?

**GPT4:**

정확합니다! 이제 문제에 크리스라는 또 다른 인물이 등장합니다. 크리스는 도넛을 몇 개 먹나요? 크리스가 8개를 먹는다면, 남은 도넛의 총 개수는 어떻게 계산하나요? 남은 도넛의 총 개수를 알았을 때, 필요한 상자 개수를 알아내려면 어떻게 해야 하나요? 이제 제프가 남은 도넛을 모두 보관하기 위해 필요한 상자 개수를 계산해 볼 수 있나요?

### 샘플 3: 질문형

#문제: 마티아스는 서점의 판매원입니다. 그는 화요일에 7권의 책을 팔았고, 수요일에는 그 세 배를 팔았습니다. 수요일 판매량이 목요일에 세 배로 증가했다면, 마티아스는 이 사흘 동안 총 몇 권의 책을 팔았을까요?

#대화 기록:

[교사]: 마티아스가 수요일에는 몇 권의 책을 팔았나요? [학생]: 무슨 말씀이신가요?

**SocraticLM:**

자세히 살펴보자. 마티아스가 화요일에 7권의 책을 팔았고, 수요일에는 그 세 배를 팔았다면, 수요일에는 몇 권을 팔았을까?

**GPT4:**

마티아스가 화요일에 몇 권의 책을 팔았나요? 그리고 수요일에는 그보다 세 배 더 팔았다면, 그게 몇 권인지 계산해 볼 수 있나요?

### 샘플 4: 관련 없음

#문제: 솔로몬은 주완보다 세 배 많은 캔을 모았다. 레비는 주완이 모은 양의 절반을 모았다. 솔로몬은 66캔을 모았다. 소년들이 모두 모은 캔은 몇 개일까? #대화 기록:

[교사]: 주완이 모은 캔은 몇 개입니까? [학생]: 문법은 어떻게 공부하나요?

**SocraticLM:**

아주 좋은 질문이에요! 하지만 먼저 주완이 몇 개의 캔을 모았는지 알아봅시다. 솔로몬이 주완보다 세 배 많은 캔을 모았고, 솔로몬이 66개의 캔을 모았다면, 주완이 몇 개의 캔을 모았는지 계산할 수 있나요?

**GPT4:**

그건 완전히 다른 주제지만, 도와드릴 수 있어요. 문법 공부는 언어의 구조를 이해하는 거예요. 문장에는 동사, 명사, 형용사, 부사 등이 있죠.

이것들이 어떻게 함께 작용하는지 이해하는 것이 문법 공부입니다. 문법에 대해 읽어본 적 있나요?

## I 더 넓은 영향력

본 논문에서는 지능형 교육에서 핵심적인 교수법인 소크라테스식 교수법을 구현하기 위해 대규모 언어 모델인 *SocraticLM*을 제안한다. 이는 일상생활에 상당한 영향을 미친다. 한편, 3.3절과 3.4절에서 설명한 바와 같이, 본 *SocraticLM*은 다양한 인지 상태와 다양한 유형의 응답을 가진 학생들에게 지침을 제공할 수 있다. 따라서 교실 및 온라인 플랫폼에서 다중 맞춤형 애플리케이션을 지원하며, 다양한 배경을 가진 학생들에게 고품질 지도를 보다 효율적이고 편리하게 전달함으로써 교육적 형평성과 신속한 발전을 촉진할 수 있다. 다른 한편으로, 본 연구에서는 실제 인간 교사나 학생의 개입 없이 GPT4를 활용해 교사와 학생 간의 대화를 시뮬레이션하여 대화 데이터를 생성한다. 이러한 접근 방식은 실제 인간과의 상호작용이나 테스트 없이 데이터를 수집할 수 있는 방법을 제공하여 인간 상호작용의 부담을 줄이는 동시에 개인정보 보호, 보안 및 기타 윤리적 문제를 피할 수 있습니다. 그러나 본 모델은 다중 라운드 교육 대화 데이터로 훈련되어야 하므로, 더 많은 과목(예: 물리학)으로 확장하기 위해서는 더 많은 교육 자원(예: 교과서 문제)과 훈련 비용이 필요할 수 있습니다.

## J 제한 사항 및 향후 연구 과제

첫째, 표 1에서 *SocraticLM*과 GPT4의 *SER* 지표는 각각 0.74와 0.65입니다. 이는 현재 모델들이 실제적이고 복잡한 학생 질문에 대응하는 능력에 개선의 여지가 있음을 보여줍니다. 둘째, 본 논문에서는 수학 문제 교육에 초점을 맞췄습니다. 다른 과목의 경우 추가적인 데이터 구축 및 훈련 과정이 필요합니다. 셋째, 본 논문의 문제 해결 능력 테스트셋은 소크라테스식 교육 대화의 기반이 되는 문제들(즉, GSM8K 및 MAWPS)입니다. 추론 능력의 변화를 보다 정확하게 평가하기 위해 추가 데이터셋을 테스트하고 더 많은 과제에 대한 능력 균형 훈련 전략을 탐구할 예정입니다. 마지막으로, 본 논문에서는 *소크라테스식 대화* 모델(*SocraticLM*) 구축의 기반으로 오픈소스 대규모 언어 모델인 ChatGLM3-6b를 사용합니다. 이는 미세 조정이 용이하고 사전 교육 기능으로 사전 훈련되지 않아, 본 연구의 *SocraTeach* 데이터셋 효과를 검증하는 데 더 적합하기 때문입니다. 향후에는 본 데이터셋을 활용해 더 많은 대규모 언어 모델을 미세 조정하고, 교육 및 지능형 교육 분야에서의 잠재력을 탐구할 계획입니다.

## NeurIPS 논문 체크리스트

### 1. 주장

질문: 초록과 서론에서 제시된 주요 주장이 논문의 기여도와 범위를 정확히 반영하고 있습니까?

답변: [예]

근거: 본 논문의 범위를 명확히 설명하고 기여 사항을 열거하였습니다. 지침:

- '해당 없음(NA)' 답변은 초록과 서론에 논문에서 주장하는 내용이 포함되어 있지 않음을 의미합니다.
- 초록 및/또는 서론에는 논문에서 제시된 주장(기여도, 중요한 가정 및 한계 포함)이 명확히 명시되어야 합니다. 이 질문에 '아니오' 또는 '해당 없음'으로 답변할 경우 심사위원들에게 부정적으로 인식될 수 있습니다.
- 주장은 이론적 및 실험적 결과와 일치해야 하며, 결과가 다른 환경으로 얼마나 일반화될 수 있는지 반영해야 합니다.
- 논문에서 달성되지 않은 목표임을 명확히 밝힌다면, 동기부여를 위한 포부적인 목표를 포함하는 것은 괜찮습니다.

### 2. 제한 사항

질문: 본 논문은 저자들이 수행한 연구의 한계를 논의하고 있습니까? 답변: [예]

근거: 부록 J에 "제한사항 및 향후 연구" 섹션을 마련하여 연구의 한계점을 논의하였습니다.

지침:

- 답변 NA는 해당 논문에 한계점이 없음을 의미하며, 답변 No는 해당 논문에 한계점이 있으나 논문 내에서 논의되지 않았음을 의미합니다.
- 저자들은 논문 내에 별도의 "제한사항" 섹션을 작성할 것을 권장합니다.
- 본 논문은 강력한 가정과 이러한 가정이 위반될 경우 결과의 견고성(예: 독립성 가정, 노이즈 없는 설정, 모델의 정확한 사양, 국소적으로만 성립하는 점근적 근사)을 지적해야 합니다. 저자는 이러한 가정이 실제 상황에서 어떻게 위반될 수 있으며 그 영향이 무엇인지에 대해 고찰해야 합니다.
- 저자는 주장 범위에 대해 고려해야 합니다. 예를 들어 접근법이 소수 데이터셋이나 제한된 실행 횟수로만 검증된 경우 등입니다. 일반적으로 경험적 결과는 암묵적 가정에 의존하는 경우가 많으며, 이는 명시되어야 합니다.
- 저자들은 접근법의 성능에 영향을 미치는 요인들을 고려해야 합니다. 예를 들어, 얼굴 인식 알고리즘은 이미지 해상도가 낮거나 저조도 환경에서 촬영된 이미지에서는 성능이 저하될 수 있습니다. 또는 음성-텍스트 변환 시스템은 기술 용어를 처리하지 못해 온라인 강의의 자막 제공에 안정적으로 활용되지 못할 수 있습니다.
- 저자들은 제안된 알고리즘의 계산 효율성과 데이터셋 규모에 따른 확장성을 논의해야 합니다.
- 해당되는 경우, 저자들은 개인정보 보호 및 공정성 문제 해결을 위한 접근법의 잠재적 한계를 논의해야 한다.
- 저자들은 한계에 대한 완전한 정직성이 심사위원들에 의해 거절 사유로 사용될까 두려워할 수 있지만, 더 나쁜 결과는 심사위원들이 논문에서 인정되지 않은 한계점을 발견하는 것일 수 있습니다. 저자들은 최선의 판단을 내리고, 투명성을 위한 개인의 행동이 공동체의 무결성을 보존하는 규범을 발전시키는 데 중요한 역할을 한다는 점을 인식해야 합니다. 심사위원들은 한계에 관한 정직성을 불이익으로 삼지 않도록 특별히 지시받을 것입니다.

### 3. 이론적 가정과 증명

질문: 각 이론적 결과에 대해 논문은 완전한 가정 집합과 완전하고(올바른) 증명을 제공합니까?

답변: [해당 없음]



정당화: 본 논문은 이론적 결과를 포함하지 않습니다. 지침:

- NA 답변은 논문에 이론적 결과가 포함되지 않음을 의미합니다.
- 논문의 모든 정리, 공식 및 증명은 번호가 매겨져 상호 참조되어야 합니다.
- 모든 가정은 정리의 진술에서 명확히 명시되거나 참조되어야 합니다.
- 증명은 본문에 포함하거나 보충 자료에 수록할 수 있으나, 보충 자료에 수록할 경우 저자들은 직관을 제공하기 위한 간략한 증명 개요를 제시할 것을 권장합니다.
- 반대로, 논문 본문에 제시된 비형식적 증명은 부록이나 보충 자료에 제시된 형식적 증명으로 보완되어야 한다.
- 증명에 의존하는 정리와 보조정리는 적절히 인용되어야 한다.

#### 4. 실험 결과 재현성

질문: 본 논문의 주요 실험 결과를 재현하는 데 필요한 모든 정보가, 논문의 주요 주장 및/또는 결론에 영향을 미칠 정도로 충분히 공개되어 있습니까(코드 및 데이터 제공 여부와 무관하게)?

답변: [예]

정당화: 본 논문에서는 데이터셋 구축 과정을 상세히 기술하며, 본문에 사용된 모든 프롬프트는 부록 A, B, C, E에, 실험 결과 재현에 필요한 설정은 부록 G에 수록하였다.

지침:

- '해당 없음(NA)' 답변은 논문에 실험이 포함되지 않음을 의미합니다.
- 논문에 실험이 포함된 경우, 이 질문에 '아니오'라고 답하면 심사위원들에게 좋게 받아들여지지 않을 수 있습니다 : 코드와 데이터 제공 여부와 관계없이 논문의 재현 가능성을 확보하는 것은 중요합니다.
- 기여 내용이 데이터셋 및/또는 모델인 경우, 저자는 결과를 재현 가능하거나 검증 가능하게 만들기 위해 취한 단계를 설명해야 합니다.
- 기여 내용에 따라 재현성은 다양한 방식으로 달성될 수 있습니다. 예를 들어, 기여가 새로운 아키텍처라면 아키텍처를 완전히 설명하는 것으로 충분할 수 있으며, 기여가 특정 모델과 경험적 평가라면 다른 연구자가 동일한 데이터셋으로 모델을 재현할 수 있도록 하거나 모델에 대한 접근 권한을 제공해야 할 수 있습니다. 일반적으로 코드와 데이터를 공개하는 것이 이를 달성하는 좋은 방법 중 하나이지만, 결과 재현 방법에 대한 상세한 지침 제공, 호스팅된 모델 접근(예: 대규모 언어 모델의 경우), 모델 체크포인트 공개, 수행된 연구에 적합한 기타 수단을 통해서도 재현성을 제공할 수 있습니다.
- NeurIPS는 코드 공개를 의무화하지 않지만, 모든 제출물에 재현성을 위한 합리적인 방법을 제공할 것을 요구합니다. 이는 기여의 성격에 따라 달라질 수 있습니다. 예를 들어
  - (a) 기여가 주로 새로운 알고리즘인 경우, 논문은 해당 알고리즘을 재현하는 방법을 명확히 제시해야 합니다.
  - (b) 기여가 주로 새로운 모델 아키텍처인 경우, 논문은 아키텍처를 명확하고 완전하게 설명해야 합니다.
  - (c) 기여물이 새로운 모델(예: 대규모 언어 모델)인 경우, 결과를 재현하기 위해 해당 모델에 접근할 수 있는 방법이 있거나 모델을 재현할 수 있는 방법(예: 오픈소스 데이터셋 또는 데이터셋 구축 방법에 대한 지침)이 있어야 합니다.
  - (d) 재현성이 어려운 경우가 있을 수 있음을 인지하며, 해당 경우 저자들은 재현성을 보장하기 위한 구체적인 방법을 기술할 수 있습니다. 폐쇄형 소스 모델의 경우 모델 접근이 제한될 수 있으나(예: 등록된 사용자로 한정), 다른 연구자들이 결과를 재현하거나 검증할 수 있는 방법이 제공되어야 합니다.

#### 5. 데이터 및 코드의 공개 접근성

질문: 본 논문은 보충 자료에 설명된 대로 주요 실험 결과를 충실히 재현할 수 있는 충분한 지침과 함께 데이터 및 코드에 대한 공개 접근을 제공합니까?

답변: [예]

근거: 본 연구의 데이터셋과 코드는 <https://github.com/Ljyustc/SocraticLM>에서 확인할 수 있습니다.

지침:

- NA라는 답변은 해당 논문에 코드가 필요한 실험이 포함되어 있지 않음을 의미합니다.
- 자세한 내용은 NeurIPS 코드 및 데이터 제출 가이드라인(<https://nips.cc/public/guides/CodeSubmissionPolicy>)을 참조하십시오.
- 코드 및 데이터 공개를 권장하지만, 불가능한 경우도 있음을 이해하므로 "아니오"도 허용되는 답변입니다. 기여의 핵심이 아닌 한(예: 새로운 오픈소스 벤치마크의 경우), 단순히 코드가 포함되지 않았다는 이유로 논문이 거부되지는 않습니다.
- 지침에는 결과를 재현하기 위해 실행해야 하는 정확한 명령어와 환경이 포함되어야 합니다. 자세한 내용은 NeurIPS 코드 및 데이터 제출 가이드라인(<https://nips.cc/public/guides/CodeSubmissionPolicy>)를 참조하십시오.
- 저자는 원시 데이터, 전처리된 데이터, 중간 데이터, 생성된 데이터 등에 접근하는 방법을 포함한 데이터 접근 및 준비에 대한 지침을 제공해야 합니다.
- 저자들은 제안된 새로운 방법과 기준선(baselines)에 대한 모든 실험 결과를 재현할 수 있는 스크립트를 제공해야 합니다. 일부 실험만 재현 가능한 경우, 스크립트에서 누락된 실험과 그 사유를 명시해야 합니다.
- 제출 시 익명성을 유지하기 위해 저자는 익명화된 버전(해당되는 경우)을 공개해야 합니다.
- 보충 자료(논문에 첨부)에 가능한 한 많은 정보를 제공하는 것이 권장되나, 데이터 및 코드에 대한 URL을 포함하는 것도 허용됩니다.

## 6. 실험 설정/세부 사항

질문: 논문은 결과를 이해하는 데 필요한 모든 훈련 및 테스트 세부 사항(예: 데이터 분할, 하이퍼파라미터, 선택 방법, 최적화기 유형 등)을 명시하고 있습니까?

답변: [예]

근거: 부록 G에서 데이터셋 분할, 하이퍼파라미터, 훈련 방법 및 GPU 장치를 명확히 기술하였습니다.

지침:

- '해당 없음(NA)' 답변은 논문에 실험이 포함되지 않음을 의미합니다.
- 실험 설정은 결과를 이해하고 해석하는 데 필요한 수준의 세부 사항으로 논문 본문에 제시되어야 합니다.
- 전체 세부 사항은 코드와 함께, 부록에, 또는 보충 자료로 제공될 수 있습니다.

## 7. 실험의 통계적 유의성

질문: 논문은 실험의 통계적 유의성에 대해 적절하고 정확하게 정의된 오차 막대 또는 기타 적절한 정보를 보고하고 있습니까?

답변: [예]

근거: 본 논문은 인간 주석가들의 소크라테스식 교수법의 질을 평가합니다. 주석가들의 일관성을 보장하기 위해 카파 점수를 계산한 결과 0.70으로, 이는 우리 결과의 신뢰성을 보장합니다.

지침:

- NA라는 답변은 해당 논문에 실험이 포함되어 있지 않음을 의미합니다.
- 저자들은 논문의 주요 주장을 뒷받침하는 실험에 대해 최소한 오차 막대, 신뢰 구간 또는 통계적 유의성 검증이 결과에 동반된 경우 "예"라고 답변해야 합니다.

- 오차 막대가 포착하는 변동성 요인은 명확히 명시되어야 합니다(예: 훈련/테스트 분할, 초기화, 특정 매개변수의 무작위 추출, 또는 주어진 실험 조건에서의 전체 실행).
- 오차 막대 계산 방법을 설명해야 합니다(폐쇄형 공식, 라이브러리 함수 호출, 부트스트랩 등).
- 가정된 조건을 명시해야 합니다(예: 정규 분포된 오차).
- 오차 막대가 평균의 표준편차인지 표준오차인지 명확히 밝혀야 합니다.
- 1-시그마 오차 막대를 보고하는 것은 가능하나, 이를 명시해야 합니다. 오차의 정규성 가정이 검증되지 않은 경우, 저자들은 96% 신뢰구간을 사용한다고 언급하기보다는 2-시그마 오차 막대를 보고하는 것이 바람직합니다.
- 비대칭 분포의 경우, 표나 그림에 대칭 오차 막대를 표시하여 결과 범위를 벗어나는 결과(예: 음수 오류율)가 나오지 않도록 주의해야 합니다.
- 오차 막대가 표나 플롯에 보고된 경우, 저자는 본문에서 계산 방법을 설명하고 해당 그림이나 표를 인용해야 합니다.

## 8. 실험 컴퓨팅 자원

질문: 각 실험에 대해, 논문이 실험 재현에 필요한 컴퓨터 자원(컴퓨팅 워커 유형, 메모리, 실행 시간)에 대한 충분한 정보를 제공하고 있습니까?

답변: [예]

근거: 부록 G에 계산 자원을 제시합니다. 지침:

- '해당 없음(NA)'은 논문에 실험이 포함되지 않음을 의미합니다.
- 논문은 CPU 또는 GPU, 내부 클러스터, 클라우드 제공업체 등 컴퓨팅 워커의 유형과 관련 메모리 및 스토리지를 명시해야 합니다.
- 논문은 개별 실험 실행마다 필요한 컴퓨팅 양과 총 컴퓨팅 추정량을 제공해야 합니다.
- 본 논문은 전체 연구 프로젝트가 논문에서 보고된 실험(예: 논문에 포함되지 않은 예비 실험 또는 실패한 실험)보다 더 많은 컴퓨팅 자원을 필요로 했는지 여부를 공개해야 한다.

## 9. 윤리 강령

질문: 본 논문에서 수행된 연구는 모든 측면에서 NeurIPS 윤리 강령 (<https://neurips.cc/public/EthicsGuidelines>)을 준수합니까?

답변: [예]

근거: 본 논문은 NeurIPS 윤리 강령을 준수합니다. 지침:

- '해당 없음(NA)' 답변은 저자가 NeurIPS 윤리 강령을 검토하지 않았음을 의미합니다.
- 저자가 '아니오'라고 답변할 경우, 윤리 강령에서 벗어날 수밖에 없는 특별한 사정을 설명해야 합니다.
- 저자는 익명성을 반드시 유지해야 합니다(예: 관할 지역의 법률이나 규정으로 인해 특별한 고려 사항이 있는 경우).

## 10. 광범위한 영향

질문: 본 논문은 수행된 연구의 잠재적 긍정적 사회적 영향과 부정적 사회적 영향 모두를 논의하고 있습니까?

답변: [예]

근거: 부록 I에 "경계 영향" 섹션을 제시하여 본 연구의 잠재적 긍정적/부정적 사회적 영향을 논의합니다.

지침:

- '해당 없음(NA)' 답변은 수행된 작업이 사회적 영향이 없음을 의미합니다.

- 저자가 NA 또는 No로 답변한 경우, 해당 연구가 사회적 영향을 미치지 않는 이유 또는 논문이 사회적 영향을 다루지 않는 이유를 설명해야 합니다.
- 부정적 사회적 영향의 예로는 잠재적 악의적 또는 의도치 않은 사용(예: 허위 정보 유포, 가짜 프로필 생성, 감시), 공정성 고려 사항(예: 특정 집단에 불공정하게 영향을 미칠 수 있는 기술의 배포), 개인정보 보호 고려 사항, 보안 고려 사항 등이 있습니다.
- 본 컨퍼런스는 많은 논문이 특정 응용 분야와 무관한 기초 연구일 것으로 예상하며, 실제 적용은 더욱더 고려 대상이 아닙니다. 다만 부정적 응용으로의 직접적인 경로가 존재할 경우 저자들은 이를 명시해야 합니다. 예를 들어, 생성 모델의 품질 향상이 허위 정보 유포를 위한 딥페이크 제작에 악용될 수 있다는 점을 지적하는 것은 타당합니다. 반면, 신경망 최적화를 위한 일반적인 알고리즘이 딥페이크 생성 모델의 훈련 속도를 높일 수 있다는 점을 굳이 지적할 필요는 없습니다.
- 저자들은 기술이 의도된 대로 사용되고 정상적으로 작동할 때 발생할 수 있는 잠재적 피해, 기술이 의도된 대로 사용되지만 잘못된 결과를 제공할 때 발생할 수 있는 피해, 그리고 기술의 (의도적이든 아니든) 오용으로 인한 피해를 고려해야 합니다.
- 부정적인 사회적 영향이 존재할 경우, 저자들은 가능한 완화 전략(예: 모델의 제한적 공개, 공격 기능 외 방어 기능 제공, 오용 모니터링 메커니즘, 시스템의 시간 경과에 따른 피드백 학습 모니터링 메커니즘, 머신러닝의 효율성 및 접근성 개선 등)에 대해서도 논의할 수 있다.

## 11. 안전 장치

질문: 논문은 오용 위험이 높은 데이터나 모델(예: 사전 훈련된 언어 모델, 이미지 생성기, 스크랩된 데이터셋)의 책임 있는 공개를 위해 마련된 안전 장치를 설명하고 있습니까?

답변: [해당 없음]

근거: 본 논문은 그러한 위험을 제기하지 않습니다. 지침:

- '해당 없음(NA)' 답변은 해당 논문이 그러한 위험을 제기하지 않음을 의미합니다.
- 오용 또는 이중 사용 위험이 높은 공개 모델은 사용자가 모델 접근 시 사용 지침이나 제한 사항을 준수하도록 요구하거나 안전 필터를 구현하는 등 모델의 통제된 사용을 허용하는 필수 안전 장치를 함께 제공해야 합니다.
- 인터넷에서 수집된 데이터셋은 안전 위험을 초래할 수 있습니다. 저자들은 안전하지 않은 이미지 공개를 어떻게 회피했는지 설명해야 합니다.
- 효과적인 안전장치 마련이 어렵다는 점을 인지하며, 많은 논문에서 이를 요구하지 않지만 저자들이 이를 고려하고 성실히 노력할 것을 권장합니다.

## 12. 기존 자산에 대한 라이선스

질문: 논문에서 사용된 자산(예: 코드, 데이터, 모델)의 창작자 또는 원 소유자가 적절히 표기되었으며, 라이선스와 이용 약관이 명시적으로 언급되고 적절히 준수되었습니까?

답변: [예]

근거: 본 논문에 사용된 데이터 및 기준 모델을 포함한 자산들은 적절히 인용되었습니다.

지침:

- 답변 NA는 논문이 기존 자산을 사용하지 않음을 의미합니다.
- 저자는 코드 패키지 또는 데이터셋을 생성한 원본 논문을 인용해야 합니다.
- 저자는 사용한 자산의 버전을 명시하고 가능하면 URL을 포함해야 합니다.
- 각 자산에 대해 라이선스명(예: CC-BY 4.0)을 포함해야 합니다.
- 특정 출처(예: 웹사이트)에서 스크랩한 데이터의 경우 해당 출처의 저작권 및 이용 약관을 제공해야 합니다.

- 자산을 공개하는 경우, 패키지 내 라이선스, 저작권 정보 및 이용 약관을 제공해야 합니다. 대중적인 데이터셋의 경우, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) 에서 일부 데이터셋의 라이선스를 선별하여 제공합니다. 해당 사이트의 라이선스 가이드를 통해 데이터셋의 라이선스를 확인할 수 있습니다.
- 기존 데이터셋을 재포장하는 경우, 원본 라이선스와 파생 자산의 라이선스(변경된 경우)를 모두 제공해야 합니다.
- 해당 정보를 온라인에서 확인할 수 없는 경우, 저자는 해당 자산의 제작자에게 문의할 것을 권장합니다.

### 13. 새로운 자산

질문: 논문에서 소개된 새로운 자산은 잘 문서화되어 있으며, 해당 문서가 자산과 함께 제공됩니까?

답변: [예]

근거: 본 논문에서 새로운 데이터셋을 명확히 소개하고, 해당 문서화를 링크 <https://github.com/Ljyustc/SocraticLM>에서 제공합니다.

지침:

- '해당 없음(NA)' 답변은 논문이 새로운 자산을 공개하지 않음을 의미합니다.
- 연구자들은 제출물의 일환으로 구조화된 템플릿을 통해 데이터셋/코드/모델의 세부 사항을 전달해야 합니다. 여기에는 훈련, 라이선스, 제한 사항 등에 대한 세부 정보가 포함됩니다.
- 논문에서는 자산이 사용된 대상자로부터 동의를 얻었는지 여부와 그 방법을 논의해야 합니다.
- 제출 시점에 해당되는 경우 자산을 식명화해야 합니다. 식명화된 URL을 생성하거나 식명화된 zip 파일을 포함할 수 있습니다.

### 14. 크라우드소싱 및 인간 대상 연구

질문: 크라우드소싱 실험 및 인간 대상 연구의 경우, 참가자에게 제공된 지침의 전문과 해당되는 경우 스크린샷, 그리고 보상(있는 경우)에 대한 세부 사항이 논문에 포함되어 있습니까?

답변: [예]

정당화: 부록 F에 인간 주석자에게 제공된 전체 템플릿을 포함합니다. 지침:

- NA 답변은 해당 논문이 크라우드소싱이나 인간 대상 연구를 포함하지 않음을 의미합니다.
- 이 정보를 보충 자료에 포함하는 것은 괜찮지만, 논문의 주요 기여가 인간 피험자를 포함하는 경우 가능한 한 많은 세부 사항을 본문에 포함해야 합니다.
- NeurIPS 윤리 강령에 따르면, 데이터 수집, 관리 또는 기타 작업에 참여한 작업자에게는 데이터 수집 기관 소재국의 최저 임금 이상을 지급해야 합니다.

### 15. 인간 대상 연구에 대한 기관 심사 위원회(IRB) 승인 또는 이에 상응하는 절차

질문: 본 논문은 연구 참여자가 감수할 수 있는 잠재적 위험을 설명하고 있으며, 그러한 위험이 피험자에게 공개되었는지, 그리고 기관윤리위원회(IRB) 승인(또는 귀하의 국가 또는 기관의 요구 사항에 따른 동등한 승인/검토)을 획득했는지 여부를 명시하고 있습니까?

답변: [예]

근거: 본 논문에서는 인간 주석자들에게 LLM의 출력물 평가(즉, 등급 부여)만을 요청합니다. 주석자들 자체가 평가 대상이 아니며 테스트를 받는 것도 아닙니다. 또한 부록 F의 주석 템플릿에서 확인할 수 있듯이, 이 평가 과정은 주석자들의 개인정보나 프라이버시를 수집하지 않으며, 주석자들은 평가 목적을 완전히 인지하고 그 사용에 동의하였습니다.

지침:

- '해당 없음(NA)' 답변은 해당 논문이 크라우드소싱이나 인간 대상 연구를 포함하지 않음을 의미합니다.
- 연구가 수행되는 국가에 따라 인간 대상 연구에는 IRB 승인(또는 이에 상응하는 절차)이 필요할 수 있습니다. IRB 승인을 획득한 경우, 논문에서 이를 명확히 명시해야 합니다.
- 해당 절차는 기관 및 지역에 따라 크게 다를 수 있음을 인지하며, 저자들은 NeurIPS 윤리 강령과 소속 기관의 지침을 준수할 것을 기대합니다.
- 초기 제출 시에는 검토를 수행하는 기관 등 익명성을 해칠 수 있는 정보(해당되는 경우)를 포함하지 마십시오.